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

Author *

July 9, 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 In preparations

Definition 1.1 is the definition of the proximal gradient operator, which is equivalent to the gradient descent operator when the non-smooth part of the objective is the zero function.

To show the convergence of a stochastic case of the Nesterov's accelerated proximal gradient, we prepared Lemma 1.10 and, 1.16. They are crucial in the derivation of the convergence. The derivation for the convergence rate of a stochastic accelerated variant of Nesterov's accelerated proximal gradient method is in the next section.

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

1.1 Basic definitions

{def:pg-opt}

Definition 1.1 (Proximal gradient operator)**.** *Suppose* $F = f + g$ *with* $\mathrm{ri}(\mathrm{dom}\, f) \cap \mathrm{ri}(\mathrm{dom}\, g) \neq \emptyset$*, and* f *is a differentiable function. Let* $\beta > 0$*. Then, we define the proximal gradient operator* T_β *as*

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

Remark 1.2. *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 1.3 (Bregman Divergence)**.** *Let* $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ *be a differentiable function. Then, for all the Bregman divergence* $D_f : \mathbb{R}^n \times \mathrm{dom}\,\nabla f \to \mathbb{R}$ *is defined as:*

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

Remark 1.4. *If,* f *is* $\mu \geq 0$ *strongly convex and* L *Lipschitz smooth then, its Bregman Divergence has for all* $x, y \in \mathbb{R}^n$*:* $\mu/2\|x - y\|^2 \leq D_f(x, y) \leq L/2\|x - y\|^2$.

1.2 Properties of functions, characterizations

{def:semi-scnvx} The definitions are ordered from the weakest to strongest.

Definition 1.5 (semi strongly convex function NEW)**.** *A function* $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ *is a semi strongly convex function, abbreviated as "Semi-SCNVX" with respective to a linear mapping* $A \in \mathbb{R}^{m\times n}$ *if* $F - \frac{1}{2}\|Ax\|^2$ *is a convex function.*

Remark 1.6. *Any* $\mu \geq 0$ *strongly convex function is Semi-SCNVX with* $A = \sqrt{\mu}I$*. But the converse is not true because a seminorm is not a norm. One feature of a Semi-SCNVX function is that it doesn't have a unique minimizer which differs it from strong convexity. It may not have a unique minimizer because it's not necessary that* $\ker A = \{\mathbf{0}\}$.

{def:seminorm-smooth-scnvx}

Definition 1.7 (semi relative smoothness and Semi-SCNVX NEW)**.** *Let* $m, n \in \mathbb{N}$ *be a natural numbers. Let* $f : \mathbb{R}^m \to \mathbb{R}$ *be a differentiable function and full domain. If there exists* $A_1 : \mathbb{R}^{m\times n}$ *and,* $A_2 \in \mathbb{R}^{m\times n}$ *matrices such that it satisfies:*

$$(\forall x \in \mathbb{R}^m)(\forall y \in \mathbb{R}^m)\ \frac{1}{2}\|A_1x - A_1y\|^2 \leq D_f(x, y) \leq \frac{1}{2}\|A_2x - A_2y\|^2.$$

Then, we call this function is semi-relative smooth with respect to A_1*, and semi convex to* A_2

Remark 1.8. *The definition exchanged the* $\|\cdot\|^2$ *for a seminorm squared:* $x \mapsto \|A_1 x\|^2$ *with some* $A : \mathbb{R}^m \to \mathbb{R}^n$ *for the definition of relative smoothness and relative strong convexity. Obviously, it has* $\|A_1 x\| \le \|A_2 x\|$ *for all* $x \in \mathbb{R}^m$*, which further implies that* $\ker A_1 \supseteq \ker A_2$.

{thm:semi-scnvx-equiv}

Theorem 1.9 (semi Jesen inequality NEW)**.** *A function* F *is Semi-SNCVX with* $A \in \mathbb{R}^{m\times n}$ *(Definition 1.5) if and only if, for all* $x, y \in \mathbb{R}^n$ *and,* $\lambda \in [0,1]$ *it satisfies the inequality:*

$$F(\lambda x + (1-\lambda)y) \le \lambda F(x) + (1-\lambda)F(y) - \frac{\lambda(1-\lambda)}{2}\|Ay - Ax\|^2.$$

Proof. For all $\lambda \in \mathbb{R}, x \in \mathbb{R}^n, y \in \mathbb{R}^n$ it has $-1/2\|A(\lambda x + (1-\lambda)y)\|^2 = (1/2)(\lambda\|Ax\|^2 + (1-\lambda)\|Ay\|^2 - \lambda(1-\lambda)\|Ax - Ay\|^2)$ by verifying:

$$\begin{aligned}
& -\frac{1}{2}\|A(\lambda x + (1-\lambda)y)\|^2 + \left(\frac{\lambda}{2}\|Ax\|^2 + \frac{1-\lambda}{2}\|Ay\|^2 - \frac{\lambda(1-\lambda)}{2}\|Ay - Ax\|^2\right) \\
&= -\frac{1}{2}\left(\lambda^2\|Ax\|^2 + (1-\lambda)^2\|Ay\|^2 - 2\lambda(1-\lambda)\langle Ax, Ay\rangle\right) \\
&\quad + \left(\frac{\lambda}{2} - \frac{\lambda(1-\lambda)}{2}\right)\|Ax\|^2 + \left(\frac{1-\lambda}{2} - \frac{\lambda(1-\lambda)}{2}\right)\|Ay\|^2 - \lambda(1-\lambda)\langle Ay, Ax\rangle \\
&= -\frac{\lambda^2}{2}\|Ax\|^2 - \frac{(1-\lambda)^2}{2}\|Ay\|^2 \\
&\quad + \left(\frac{\lambda}{2} - \frac{\lambda-\lambda^2}{2}\right)\|Ax\|^2 + \left(\frac{1-\lambda}{2} - \frac{\lambda-\lambda^2}{2}\right)\|Ay\|^2 \\
&= 0
\end{aligned}$$

Using the above result we can prove the equivalency because

$$\begin{aligned}
0 &\le F(\lambda x + (1-\lambda)y) + \lambda F(x) + (1-\lambda)F(y) - \frac{\lambda(1-\lambda)}{2}\|Ay - Ax\|^2 \\
&= F(\lambda x + (1-\lambda)y) - \frac{1}{2}\|A(\lambda x + (1-\lambda)y)\|^2 + \lambda F(x) - \frac{\lambda}{2}\|Ax\|^2 + (1-\lambda)F(y) - \frac{1-\lambda}{2}\|Ay\|^2 \\
&\quad - \frac{\lambda(1-\lambda)}{2}\|Ay - Ax\|^2 + \frac{1}{2}\|A(\lambda x + (1-\lambda)y)\|^2 + \frac{1}{2}\|Ax\|^2 + \frac{1}{2}\|Ay\|^2 \\
&= F(\lambda x + (1-\lambda)y) - \frac{1}{2}\|A(\lambda x + (1-\lambda)y)\|^2 \\
&\quad + \lambda\left(F(x) - \frac{1}{2}\|Ax\|^2\right) + (1-\lambda)\left(F(y) - \frac{1}{2}\|Ay\|^2\right).
\end{aligned}$$

The last line shows that the function $F(x) - \frac{1}{2}\|Ax\|^2$ is convex, the chain of equality shows the equivalence. □

{thm:jesen}

Theorem 1.10 (Jensen's inequality)**.** *Let* $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ *be a* $\mu \ge 0$ *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 1.11. *If x, y is out of $\operatorname{dom} F$, the inequality still work by convexity.*

{thm:smooth-aff-sq-scnvs-fxn} The following theorem classifies a class of semi strongly convex function.

Theorem 1.12 (affine composition with strong convexity and smoothness)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L smooth and, $\mu \geq 0$ strongly convex. Let $A : \mathbb{R}^m \to \mathbb{R}^n$, $b \in \mathbb{R}^n$ be arbitrary. Let $h : \mathbb{R}^m \to \mathbb{R} = x \mapsto f(Ax - b)$, then the function h satisfies:*

$$(\forall x \in \mathbb{R}^m)(\forall y \in \mathbb{R}^m)\ \frac{\mu}{2}\|Ax - Ay\|^2 \leq D_h(x,y) \leq \frac{L}{2}\|Ax - Ay\|^2.$$

Therefore, it satisfies Definition 1.5 with $A_1 = \sqrt{L}A, A_2 = \sqrt{\mu}A$.

Proof. Then the Bregman divergence of h is:

$$\begin{aligned} D_h(x,y) &= h(x) - h(y) - \langle \nabla h(y), x - y\rangle \\ &= f(Ax-b) - f(Ay-b) - \langle A^T \nabla f(Ay-b), x - y\rangle \\ &= f(Ax-b) - f(Ay-b) - \langle \nabla f(Ay-b), Ax - Ay\rangle \\ &= f(Ax-b) - f(Ay-b) - \langle \nabla f(Ay-b), Ax - b - (Ay - b)\rangle \\ &= D_f(Ax-b, Ay-b). \end{aligned}$$

Since f is L smooth and $\mu \geq 0$ strongly convex, it means

$$\begin{aligned} \frac{\mu}{2}\|Ax - Ay\|^2 &= \frac{\mu}{2}\|Ax - b - (Ay - b)\|^2 \\ &\leq D_f(Ax-b, Ay-b) \\ &= D_h(x,y) \\ &\leq \frac{L}{2}\|Ax - Ay\|^2. \end{aligned}$$

$\square$

The following definition defines the concept of relative smoothness. We build the proximal gradient inequality for the class of Semi-SCNVX functions.

1.3 Important inequalities

The following assumptions are ordered from weakest to strongest. All of them are some of
{ass:snorm-smth-p-nsmth} smooth with a non-smooth function.

Assumption 1.13 (affine S-CNVX composition plus non-smooth)**.** *Let $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^n$ Let $F(x) : \mathbb{R}^m \mapsto \overline{\mathbb{R}} := x \mapsto f(x) + g(x)$. Assume that:*

(i) *$f : \mathbb{R}^m \to \mathbb{R} := x \mapsto h(Ax - b)$ where $h : \mathbb{R}^n \to \mathbb{R}$ is L Lipschitz smooth and, $\mu \geq 0$ strongly convex, then it satisfies Theorem 1.12 with $L > \mu \geq 0$ and $A \in \mathbb{R}^{m\times n}$.*

{ass:smooth-plus-nonsmooth} (ii) *$g : \mathbb{R}^m \mapsto \overline{\mathbb{R}}$ is a convex, proper and closed function.*

Assumption 1.14 (smooth add nonsmooth)**.** *The function $F = f + g$ where $f : \mathbb{R}^n \to \mathbb{R}$ is an L Lipschitz smooth and $\mu \geq 0$ strongly convex function. The function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a*
{ass:smooth-plus-nonsmooth-x} *closed convex proper function.*

Assumption 1.15 (admitting minimizers)**.** *Let $F = f + g$ satisfies 1.14 and in addition assume that the set of minimizers $X^+ := \operatorname*{argmin}_x F(x)$ is non-empty.*
{thm:pg-ineq}

Theorem 1.16 (proximal gradient inequality)**.** *Let function F satisfies Assumption 1.14, so it's $\mu \geq 0$ strongly convex. For any $x \in \mathbb{R}^n$, define $x^+ = T_L(x)$. Then, there exists a $B \geq 0$ such that $D_f(x^+, x) \leq B/2\|x^+ - x\|^2$ and, for all $z \in \mathbb{R}^n$ it satisfies the inequality:*

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

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 1.17. *The theorem is the same as in Nesterov's book [4, 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 2.6 with line search routine to determine L_k. 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 [3].*

{thm:pg-ineq-semi-scnvx} The following theorem attempts to generalize Theorem 1.16.

Theorem 1.18 (proximal gradient inequality with semi-scnvx NEW)**.** *Suppose that $F : \mathbb{R}^m \to \overline{\mathbb{R}} := x \mapsto f(x) + g(x)$ satisfies Assumption 1.13 with $L > \mu \geq 0$ and $A \in \mathbb{R}^{m\times n}$. For* *any $x \in \mathbb{R}^n$, let $x^+ = T_B(x|F)$. Let $\sigma_{\min}(A)$ denotes the minimum non-zero singular value of A in absolute value. Let $\Pi = \Pi_{\ker A}$ be the linear operator that project onto the kernel of A. Then, there exists some $B \geq 0$ such that $D_f(x^+, x) \leq \frac{B}{2}\|x - x^+\|^2$ and, for all $z \in \mathbb{R}^m$ it satisfies the inequality:*

$$
\begin{aligned}
0 &\leq F(z) - F(x^+) - \frac{\mu}{2}\|Az - Ax\|^2 + \frac{B}{2}\|z - x\|^2 - \frac{B}{2}\|z - x^+\|^2 \\
&\leq F(z) - F(x^+) - \frac{B}{2}\|z - x^+\|^2 + \frac{B - \sigma_{\min}(A)^2\mu}{2}\|(I - \Pi)(z - x)\|^2 \\
&\quad + \frac{B}{2}\|\Pi(z - x)\|^2.
\end{aligned}
$$

Proof. Firstly, such a $B > 0$ exists, for example $B = L\|A\|^2$ would be an option because from Definition 1.7, it for all x, y, $D_f(x, y) \leq L/2\|Ax - Ay\|^2 \leq L/2\|A\|^2\|x - y\|^2$. But it can be much smaller.

The function $z \mapsto g(z) + f(x) + \langle \nabla f(x), z - x\rangle + \frac{B}{2}\|z - x\|^2$ inside the proximal gradient operator has the minimizer x^+. This function is also the sum of a convex, proper closed function g and, a simple quadratic and, it's $B > 0$ strongly convex hence, it satisfies the quadratic growth conditions over its minimizer $x^+ = T_B(x|F)$ so, it follows that for all $z \in \mathbb{R}^m$:

$$
\begin{aligned}
0 &\leq -\frac{B}{2}\|z - x^+\|^2 + g(z) + f(x) + \langle \nabla f(x), z - x\rangle + \frac{B}{2}\|z - x\|^2 \\
&\quad - g(x^+) - f(x) - \langle \nabla f(x), x^+ - x\rangle - \frac{B}{2}\|x^+ - x\|^2 \\
&= -\frac{B}{2}\|z - x^+\|^2 + \left(g(z) + f(z) - f(z) + f(x) + \langle \nabla f(x), z - x\rangle + \frac{B}{2}\|z - x\|^2\right) \\
&\quad + \left(-g(x^+) - f(x^+) + f(x^+) - f(x) - \langle \nabla f(x), x^+ - x\rangle - \frac{B}{2}\|x^+ - x\|^2\right) \\
&= -\frac{B}{2}\|z - x^+\|^2 + \left(F(z) - D_f(z, x) + \frac{B}{2}\|z - x\|^2\right) \\
&\quad + \left(-F(x^+) + D_f(x^+, x) - \frac{B}{2}\|x^+ - x\|^2\right) \\
&\underset{\text{(a)}}{\leq} -\frac{B}{2}\|z - x^+\|^2 + \left(F(z) - D_f(z, x) + \frac{B}{2}\|z - x\|^2\right) - F(x^+) \\
&\underset{\text{(b)}}{\leq} -\frac{B}{2}\|z - x^+\|^2 + F(z) - \frac{\mu}{2}\|Az - Ax\|^2 + \frac{B}{2}\|z - x\|^2 - F(x^+) \\
&= F(z) - F(x^+) - \frac{\mu}{2}\|Az - Ax\|^2 + \frac{B}{2}\|z - x\|^2 - \frac{B}{2}\|z - x^+\|^2.
\end{aligned}
$$

At (a), we used the fact that line search asserted the condition $D_f(x^+, x) \le \frac{B}{2}\|x^+ - x\|^2$. At (b), we used the fact that f satisfies Definition 1.7 so, it has for all $x, z \in \mathbb{R}^m$, $D_f(z, x) \ge \frac{\mu}{2}\|z - x\|^2$.

Since $\Pi = \Pi_{\ker A}$ to be the projection onto the kernel of matrix A, $z - x$ can be written as $\Pi(z - x) + (z - x) - \Pi(z - x)$ and so:

$$
\begin{aligned}
&\Pi(z - x) \perp (z - x) - \Pi(z - x) \\
\implies &\|z - x\|^2 = \|\Pi(z - x)\|^2 + \|(z - x) - \Pi(z - x)\|^2
\end{aligned}
$$

Let $\sigma_{\min}(A)$ denotes the minimum non-zero singular value of A in absolute value, then it has:

$$
\begin{aligned}
\|A(x - z)\|^2 &= \|A(\Pi(z - x) + (z - x) - \Pi(z - x))\|^2 \\
&= \|A((z - x) - \Pi(z - x))\|^2 \\
&\ge \sigma_{\min}(A)^2\|(z - x) - \Pi(z - x)\|^2.
\end{aligned}
$$

Using two of the above results to keep simplifying the inequality, it yields:

$$
\begin{aligned}
0 &\le F(z) - F(x^+) - \frac{B}{2}\|z - x^+\|^2 - \frac{\mu}{2}\|Az - Ax\|^2 + \frac{B}{2}\|z - x\|^2 \\
&\le F(z) - F(x^+) - \frac{B}{2}\|z - x^+\|^2 + \frac{B - \sigma_{\min}(A)^2\mu}{2}\|(z - x) - \Pi(z - x)\|^2 \\
&\quad + \frac{B}{2}\|\Pi(z - x)\|^2. \\
&= F(z) - F(x^+) - \frac{B}{2}\|z - x^+\|^2 + \frac{B - \sigma_{\min}(A)^2\mu}{2}\|(I - \Pi)(z - x)\|^2 \\
&\quad + \frac{B}{2}\|\Pi(z - x)\|^2.
\end{aligned}
$$

Last equality we used the fact that Π is an linear operator as well. □

Remark 1.19. *When* $\ker A = \{\mathbf{0}\}$, *this theorem is equivalent to Theorem 1.16 but with* μ *being* $\sigma_{\min}(A)$ *instead.*

{thm:smnrm-jnsn-smth-nsmth}

The following theorem attempts to generalize Theorem 1.9 for relatively smooth plus non-smooth function.

Theorem 1.20 (seminorm smooth plus non-smooth Jensen NEW)**.** *Suppose that* $F : \mathbb{R}^m \to \overline{\mathbb{R}} := x \mapsto f(x) + g(x)$ *satisfies Assumption 1.13 with* $L > \mu \ge 0$ *and* $A \in \mathbb{R}^{m\times n}$. *Let* $\Pi = \Pi_{\ker A}$ *be the projection onto the kernel of* A. *Let* $\sigma_{\min}(A)$ *denote the smallest non-zero singular value of* A *in absolute value. Then, for all* $x, y \in \mathbb{R}^m$ *and,* $\lambda \in [0, 1]$ *it satisfies the inequality:*

$$
F(\lambda z + (1 - \lambda)y) \le \lambda F(z) + (1 - \lambda)F(x) - \frac{\sigma_{\min}(A)^2\lambda(1 - \lambda)\mu}{2}\|(I - \Pi)(x - y)\|^2.
$$

NOT YET FINISHED

Proof. f satisfies Definition 1.7 with $L > \mu, A$, so for all $x, y \in \mathbb{R}^m$ it has

$$0 \le D_f(x,y) - \frac{\mu}{2}\|Ax - Ay\|^2 \le \frac{L-\mu}{2}\|Ax - Ay\|^2.$$

Using some algebra (or equivalent some properties of Bregman divergence), it shows that the function $f - \mu/2\|A(\cdot)\|^2$ is a convex function, therefore, $f + g - \mu/2\|A(\cdot)\|^2 = F - \frac{\mu}{2}\|A(\cdot)\|^2 = F - \frac{1}{2}\|\sqrt{\mu}A(\cdot)\|^2$ is also a convex function. Applying Theorem 1.9 it has for all $z, x \in \mathbb{R}^m$ and, $\lambda \in [0,1]$ the inequality:

$$F(\lambda z + (1-\lambda)y) \le \lambda F(z) + (1-\lambda)F(x) - \frac{\lambda(1-\lambda)\mu}{2}\|Ax - Ay\|^2.$$

Let $\Pi = \Pi_{\ker A}$ be the projection onto the kernel of A, then it has:

$$\begin{aligned}\|Ax - Ay\|^2 &= \|A\Pi(x-y) + (I-\Pi)(x-y)\|^2 \\ &= \|A\Pi(x-y)\|^2 + \|A(I-\Pi)(x-y)\|^2 \\ &\ge 0 + \sigma_{\min}(A)^2\|(I-\Pi)(x-y)\|^2\end{aligned}$$

Therefore, it gives the weaker inequality:

$$F(\lambda z + (1-\lambda)y) \le \lambda F(z) + (1-\lambda)F(x) - \frac{\sigma_{\min}(A)^2\lambda(1-\lambda)\mu}{2}\|(I-\Pi)(x-y)\|^2.$$

□

2 Stochastic accelerated proximal gradient

In this section, we define the Stochastic accelerated Nesterov's Acceleration algorithm and, progressively build up the results to the one step convergence of the algorithm. 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-aff-comp}

Assumption 2.1 (sum of many affine composite)**.** *Let $A^{(1)}, A^{(2)}, \ldots A^{(n)}$ be a list of $\mathbb{R}^{n\times m}$ matrices. Suppose $F : \mathbb{R}^n \to \overline{\mathbb{R}} := (1/n)\left(\sum_{i=1}^n f_i(x) + g_i(x)\right)$. Assume*

(i)

{ass:sum-of-many}

Assumption 2.2 (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* $i = 1, \ldots, n$*, 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,* $g_i : \mathbb{R}^n \to \overline{\mathbb{R}}$ *is a closed convex proper function.*

Consequently, the function f *can be written as* $F = g + f$ *with* $f = (1/n)\sum_{i=1}^{n} f_i, g = (1/n)\sum_{i=1}^{n} g_i$ *therefore, it also satisfies Assumption 1.14 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 1.14. 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 2.3 (interpolation hypothesis)**.** *Suppose that* $F := (1/n)\sum_{i=1}^{n} F_i$ *satisfying Assumption 2.2. 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* $0 = f_i(\bar{x})$*.*

Consequently, each of the F_i *satisfies Assumption 1.15 with* X_i *being the set of minimizers and, under interpolation hypothesis this equates to non-empty intersections between all* X_i*, i.e:* $\bigcap_{i=1}^{n} X_i \neq \emptyset$*.*

What is the weakest possible sequence one can use for the accelerated proximal gradient based algorithm that utilizes a strong convexity constant? If we were to use the developed convergence framework for Nesterov's accelerated proximal gradient, negative momentum and, negative convergence (lower bound instead of upper bound) should be prohibited, and it means that the sequence $(\alpha_k)_{k\geq 0}$ which is going to appear in the proposed algorithm (See Definition 2.6) must satisfy the condition $\alpha_k \in (0, 1]$ for all $k \geq 0$. The following lemma with a blunt name should clarify the sufficient conditions required for the sequence to make sense.

{lemma:snapg-v2-seq-range}

Lemma 2.4 (weakest possible momentum sequence that makes sense NEW)**.** *Suppose that* $(L_k)_{k\geq 0}$ *is a sequence such that* $L_k > 0$ *for all* $k \geq 0$*. Suppose that* $(\tilde{\mu}_k)_{k\geq 0}$ *is another non-negative sequence. 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:*

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

And, the following items are true:

(i) *The expression of α_k based on previous α_{k-1} is given by:*

$$\alpha_k = \frac{L_{k-1}}{2L_k}\left(-\alpha_{k-1}^2 + \frac{\tilde{\mu}_k}{L_{k-1}} + \sqrt{\left(\alpha_{k-1} - \frac{\tilde{\mu}_k}{L_{k-1}}\right)^2 + \frac{4\alpha_{k-1}^2 L_k}{L_{k-1}}}\right) \geq 0.$$

(ii) *If, in addition, the sequence $\tilde{\mu}_k$ satisfies for all $k \geq 1$, $\frac{\tilde{\mu}_k}{L_{k-1}} < L_{k-1}/L_k$, then the sequence strictly less than one and, for all $k \geq 1$: $\alpha_k \in (0, 1)$.*

Proof. For all $k \geq 1$, re-arranging the equality it comes to solving the following equality:

$$\begin{aligned}
0 &= L_k\alpha_k^2 - \tilde{\mu}_k\alpha_k + L_{k-1}\alpha_{k-1}^2\alpha_k - L_{k-1}\alpha_{k-1}^2 \\
&= L_k\alpha_k^2 + (L_{k-1}\alpha_{k-1}^2 - \tilde{\mu}_k)\alpha_k - L_{k-1}\alpha_{k-1}^2 \\
\iff 0 &= \alpha_k^2 + L_k^{-1}(L_{k-1}\alpha_{k-1}^2 - \tilde{\mu}_k)\alpha_k - L_k^{-1}L_{k-1}\alpha_{k-1}^2 \\
\iff \alpha_k &= \frac{1}{2}\left(-L_k^{-1}(L_{k-1}\alpha_{k-1}^2 - \tilde{\mu}_k) + \sqrt{L_k^{-2}(L_{k-1}\alpha_{k-1}^2 - \tilde{\mu}_k)^2 + 4L_k^{-1}L_{k-1}\alpha_{k-1}^2}\right) \\
&= \frac{L_{k-1}}{2L_k}\left(-\alpha_{k-1}^2 + \frac{\tilde{\mu}_k}{L_{k-1}} + \sqrt{\left(\alpha_{k-1}^2 - \frac{\tilde{\mu}_k}{L_{k-1}}\right)^2 + \frac{4L_k}{L_{k-1}}\alpha_{k-1}^2}\right)
\end{aligned}$$

Here, we take the positive root of the quadratic so that it ensures $\alpha_k \geq 0$. This is true by induction. If $\alpha_{k-1} \geq 0$ then the $\frac{4L_k}{L_{k-1}}\alpha_{k-1}^2 \geq 0$ hence, the square root is greater than the term outside it so, $\alpha_k \geq 0$ too.

Assume inductively that $\alpha_{k-1} \geq 0$. Next, we want to find the conditions needed such that

$\alpha_k < 1$. To start, we complete the square root inside the square root:

$$
\begin{aligned}
0 &\le \left(\alpha_{k-1}^2 - \frac{\tilde{\mu}_k}{L_{k-1}}\right)^2 + \frac{4L_k}{L_{k-1}}\alpha_{k-1}^2 \\
&= \alpha_{k-1}^4 + \left(\frac{\tilde{\mu}_k}{L_{k-1}}\right)^2 - 2\alpha_{k-1}^2\frac{\tilde{\mu}_k}{L_{k-1}} + \frac{4L_k}{L_{k-1}}\alpha_{k-1}^2 \\
&= \alpha_{k-1}^4 + \left(\frac{\tilde{\mu}_k}{L_{k-1}}\right)^2 + \alpha_{k-1}^2\left(\frac{-2\tilde{\mu}_k}{L_{k-1}} + \frac{4L_k}{L_{k-1}}.\right) \\
&= \alpha_{k-1}^4 + \left(\frac{\tilde{\mu}_k}{L_{k-1}}\right)^2 + \alpha_{k-1}^2\left(\frac{4L_k - 2\tilde{\mu}_k}{L_{k-1}}\right) \\
&= \alpha_{k-1}^4 + \alpha_{k-1}^2\left(\frac{4L_k - 2\tilde{\mu}_k}{L_{k-1}}\right) + \left(\frac{2L_k - \tilde{\mu}_k}{L_{k-1}}\right)^2 - \left(\frac{2L_k - \tilde{\mu}_k}{L_{k-1}}\right)^2 + \left(\frac{\tilde{\mu}_k}{L_{k-1}}\right)^2 \\
&= \left(\alpha_{k-1}^2 + \frac{2L_k - \tilde{\mu}_k}{L_{k-1}}\right)^2 - \left(\frac{2L_k - \tilde{\mu}_k}{L_{k-1}}\right)^2 + \left(\frac{\tilde{\mu}_k}{L_{k-1}}\right)^2 \\
&= \left(\alpha_{k-1}^2 + \frac{2L_k - \tilde{\mu}_k}{L_{k-1}}\right)^2 + \frac{\tilde{\mu}_k^2 - 4L_k^2 - \tilde{\mu}_k^2 + 4L_k\tilde{\mu}_k}{L_{k-1}^2} \\
&= \left(\alpha_{k-1}^2 + \frac{2L_k - \tilde{\mu}_k}{L_{k-1}}\right)^2 + \frac{4L_k\tilde{\mu}_k - 4L_k^2}{L_{k-1}^2} \\
&= \left(\alpha_{k-1}^2 + \frac{2L_k - \tilde{\mu}_k}{L_{k-1}}\right)^2 + 4\left(\frac{L_k}{L_{k-1}} \cdot \frac{\tilde{\mu}_k}{L_{k-1}} - 1\right) \\
&< \left(\alpha_{k-1}^2 + \frac{2L_k - \tilde{\mu}_k}{L_{k-1}}\right)^2.
\end{aligned}
$$

On the last inequality, we used our assumption that the sequence $\tilde{\mu}_k, L_k$ satisfies $\frac{\tilde{\mu}_k}{L_{k-1}} < \frac{L_{k-1}}{L_k}$. Substitute it back into the expression previous obtained for α_k, using the monotone property of the function $\sqrt{\cdot}$, it gives the inequality

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

□

Remark 2.5. *Let's do some sanity check for the lemma we just derived. The sequence L_k will be from the Lipschitz line search routine of the accelerated proximal gradient method.*

(i) *Let's assume the obvious choice of $L_k = \max_{i=1,\dots,n} K^{(i)}$ for all $k = 1, 2, \dots$ given an objective function F satisfying Assumption 2.2. Then, the sufficient condition for the* *second item translates to $\tilde{\mu}_i/L_k < 1$. Hence, if we choose $\tilde{\mu}_i$ to be a constant sequence of 0 then it works out to have $\alpha_k \in (0,1)$ for all $k = 1, 2, \ldots$.*

If F has $L \geq \mu$ so, the function is non-trivial, then choose $\tilde{\mu}_i = \mu$, the true strong convexity parameter then it also works out.

(ii) *Let's assume that some type of monotone line search routine is used for the algorithm making $L_0 \leq L_1 \leq \ldots \leq L_k \leq \ldots$ to be a non-decreasing sequence, then it requires $\tilde{\mu}_k/L_{k-1} \leq L_{k-1}/L_k$.*

Well, it will still make sense because one such choice could be $\tilde{\mu}_k = \rho \min_{i=1,\ldots,k} L_{i-1}/L_i$ for some $\rho \in (0,1)$.

{def:snapg-v2}

Definition 2.6 (SNAPG-V2)**.** *Let F satisfies Assumption 2.2. 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 $v_{-1} = x_{-1}, \alpha_0 = 1$. Let $\tilde{\mu} \geq 0$ be a constant that is fixed. The SNAPG generates the sequence $(y_k, x_k, v_k)_{k\geq 0}$ such that for all $k \geq 0$ they satisfy:*

$$
\begin{aligned}
&\alpha_k \in (0,1) : (L_{k-1}/L_k)(1-\alpha_k)\alpha_{k-1}^2 = \alpha_k\left(\alpha_k - \tilde{\mu}/L_k\right),\\
&\tau_k = L_k(1-\alpha_k)\left(L_k\alpha_k - \mu^{(I_k)}\right)^{-1},\\
&y_k = (1+\tau_k)^{-1}v_{k-1} + \tau_k(1+\tau_k)^{-1}x_{k-1},\\
&L_k > 0 : D_f(x_k, y_k) \leq L_k/2\|y_k - x_k\|^2,\\
&x_k = T_{L_k}(y_k|F_{I_k}),\\
&v_k = x_{k-1} + \alpha_k^{-1}(x_k - x_{k-1}).
\end{aligned}
$$

Remark 2.7. *$\tilde{\mu}_k, L_k$ are not necessary a random variable because they are determined by a line-search like conditions, consequently $(\alpha_k)_{k\geq 0}$, whether they are a random variable depends on the line search procedures. Otherwise, all the iterates (x_k, y_k, z_k) are random variable determined by I_k when conditioned on all previous $I_{k-1}, I_{k-2}, \ldots, I_0$.*

NEW. One may notice that α_k requires L_k which comes before L_k, x_k which are needed in advanced for α_k. This may seem off since no algorithm can know what L_k to choose in advanced to determine the line search. But, it is important to note that in here, we defined a sequence of conditions on the iterates x_k, y_k, z_k, and auxiliary sequences α_k, L_k which is not a definition of any algorithm. It is quantifying the conditions needed for an algorithm that actually implements it.

For the trivial case where we don't need to worry about it is when $L_k = \max_{i=1,\ldots,n} K^{(i)}$. See Chambolle, Calatroni [2] for an implementation of linear search with backtracking for the FISTA algorithm, it is how one would implement it in the deterministic case.

The following lemma state the relationships of the iterates generated by SNAPG-V2. They are needed for the convergence proof.

{lemma:snapg2-itrs-props}

Lemma 2.8 (properties of the iterates NEW)**.** *Suppose that the iterates $(z_k, x_k, y_k)_{k\geq 0}$ and sequence $(\alpha_k)_{k\geq 1}$ are produced by an algorithm satisfying Definition 2.6. Let $\bar{x} \in \mathbb{R}^n$. Define the sequence $z_k = \alpha_k \bar{x} + (1-\alpha_k)x_{k-1}$. Then, the following are true:*

{lemma:snapg2-itrs-props-item1}

(i) *For all $k \geq 1$ it has:*

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

{lemma:snapg2-itrs-props-item2}

(ii) *For all $k \geq 1$, it has: $z_k - x_k = \alpha_k(x - \bar{x})$*

Proof. **Proof of (i)**. From Definition 2.6, 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 (ii). From Definition 2.6 it has directly:

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

{lemma:snagp2-one-step-s1-proto} □

Lemma 2.9 (SNAPG-V2 one step convergence prototype stage-I)**.**

{thm:snapg2-one-step} *Proof.* □

Theorem 2.10 (SNAPG-V2 one step convergence)**.** *Let F satisfies assumption 2.3. Suppose that an algorithm satisfying Definition 2.6 uses this F. Let $\mathbb{E}_k$ denotes the expectation conditioned on $I_0, I_1, \ldots, I_{k-1}$. Then, for all $k \ge 1$, it has the following inequality*

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

And for $k = 0$, it has

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

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. The proof is long so, we use letters and subscript under relations such as $\underset{(\cdot)}{=}, \underset{(\cdot)}{\geq}$ to indicate which result is used going from the previous expression to the next. We list the following intermediate results, (d)-(g) are proved at the end of the proof.

(a) We can use proximal gradient inequality from Theorem 1.16 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 1.10 with $z = z_k$ on F_i.

(c) The sequence $(\alpha_k)_{k\geq 0}$ has $(L_{k-1}/L_k)(1-\alpha_k)\alpha_{k-1}^2 = \alpha_k\left(\alpha_k - \mu/L_k\right)$.

(d) Prove in Lemma 2.8 (i) we use the 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 Lemma 2.8 (ii), we use: $(\forall k \geq 1)\ z_k - x_k = \alpha_k(\bar{x} - v_k)$.

(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 $k \geq 1$:

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

(h) Because we assumed interpolation hypothesis in Assumption 2.3, 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}

$$\begin{aligned}
0 &\leq F_i(z_k) - F_i(x_k) - \frac{L_k}{2}\|z_k - x_k\|^2 + \frac{L_k - \mu^{(i)}}{2}\|z_k - y_k\|^2 \\
&\underset{\text{(b)}}{\leq} \alpha_k F_i(\bar{x}) + (1-\alpha_k)F_i(x_{k-1}) - F_i(x_k) \\
&\quad - \frac{\mu^{(i)}\alpha_k(1-\alpha_k)}{2}\|\bar{x} - x_{k-1}\|^2 - \frac{L_k}{2}\|z_k - x_k\|^2 + \frac{L_k - \mu^{(i)}}{2}\|z_k - y_k\|^2.
\end{aligned} \tag{2.1}$$

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 \\
\underset{(d)}{=} & -\frac{\mu^{(i)}\alpha_k(1-\alpha_k)}{2}\|\bar{x}-x_{k-1}\|^2 \\
& +\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 \\
& +\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 \\
& +\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 \\
& +\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 \\
& +\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{(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 \\
& +\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 \\
& +\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{(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 \\
& +\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(\tilde{\mu}-\mu^{(i)})}{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+\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) \\
& +\frac{\alpha_k(\tilde{\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
\end{aligned}
$$

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

Substituting the above back to the tail of Inequality (2.1) it gives:

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

Recall that $i = I_k$ is the random variable from Definition 2.6. Rearranging the last expression in the above equality chain can be conveniently written as

{ineq:snapg2-one-step-presult1}

$$
\begin{aligned}
&F_{I_k}(x_k) - F_{I_k}(\bar{x}) + \frac{L_k\alpha_k^2}{2}\|\bar{x} - v_k\|^2 \\
&\le (1-\alpha_k)\left(F_{I_k}(x_{k-1}) - F_{I_k}(\bar{x}) + \frac{\alpha_{k-1}^2 L_{k-1}}{2}\|v_{k-1} - \bar{x}\|^2\right) \\
&\quad + \frac{\alpha_k(\tilde{\mu} - \mu^{(I_k)})}{2}\|\bar{x} - v_{k-1}\|^2 + \frac{(\alpha_k - 1)\mu^{(I_k)}\left(L_k\alpha_k - \mu^{(I_k)}\right)}{2\left(L_k - \mu^{(I_k)}\right)}\|x_{k-1} - v_{k-1}\|^2.
\end{aligned} \tag{2.2}
$$

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 (2.2) 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 (2.2), using the linearity property while taking the conditional expectation yields:

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

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 2.2) 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 (2.2), we have the one-step inequality in expectation:

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

Finally, we show the base case. When $k = 0$, by assumption it had $\alpha_0 = 1$ hence τ_0 in Definition 2.6 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 1.16 with $z = z_0$ and, assume a successful line search with L_0, it yields:

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

Re-arranging and taking the expectation it yields:

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

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:

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

□

3 Convergence rate of the algorithm under various circumstances

The previous section highlighted a generic convergence results from one iteration of the algorithm, however, there are a lot of loose ends. This section will deal with those.

4 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 around it.

(ii) Roll up the inequality in Theorem 2.10 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 2.3 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] L. Calatroni and A. Chambolle, *Backtracking strategies for accelerated descent methods with smooth composite objectives*, SIAM Journal on Optimization, 29 (2019), pp. 1772–1798.

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

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