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

Author *

July 7, 2025

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

Abstract

This file is for communication purposes between collaborators.

1 Nesterov's Accelerated Gradient

1.1 In preparations

Assumption 1.1 states the standard assumption of smooth plus non-smooth structured optimization problems for proximal gradient based algorithm. Assumption 1.2 adds the assumption for the existence of minimizers. Definition 1.3 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.

Finally, to show the convergence of a stochastic case of the Nesterov's accelerated proximal gradient, we prepared Lemma 1.9 and, 1.7. They are crucial in the derivation of the

---

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

convergence. The derivation for the convergence rate of a stochastic accelerated variant of Nester's accelerated proximal gradient method is in the next section.

Assumption 1.1 (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 closed convex proper function.*

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

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 $x, y \in \mathbb{R}^n$: $\mu/2\|x - y\|^2 \leq D_f(x, y) \leq L/2\|x - y\|^2$.*

Theorem 1.7 (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:*

$$0 \leq F(z) - F(x^+) - \frac{B}{2}\|z - x^+\|^2 + \frac{B - \mu}{2}\|z - x\|^2$$

$$= F(z) - F(x^+) - \langle B(x - x^+), z - x \rangle - \frac{\mu}{2}\|z - x\|^2 - \frac{B}{2}\|x - x^+\|^2.$$

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 1.8. *The theorem is the same as in Nesterov's book [**?**, 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 [**?**, Theorem 10.16]. The only generalization here is parameter B which made to accommodate algorithm that implements Definition 1.17 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 [**?**].*

2

Theorem 1.9 (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 1.10. *If x, y is out of $\mathrm{dom}\, F$, the inequality still work by convexity.*

Definition 1.11 (semi quasi strong convexity (NEW)). *A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is semi quasi strongly convex if, there exists a constant $\mu \geq 0$ and, a non-zero linear mapping $A : \mathbb{R}^n \to \mathbb{R}^m$ such that for all $x \in \mathbb{R}^n, y \in RR^n$ and $\lambda \in [0,1]$ it satisfies:*

$$F(\lambda x + (1-\lambda)y) \leq \lambda F(x) + (1-\lambda)F(y) - \frac{\mu \lambda (1-\lambda)}{2} \|A(y-x)\|^2.$$

Remark 1.12. *It is strong convexity but instead of being strongly convex to the function $\mu/2 \|\cdot\|^2$, it's $x \mapsto \mu/2 \|Ax\|^2$ instead. This is an example of uniform convexity because it also fits in Bauchke [1, Definition 10.7], with $\phi = \|A\|^2 \|x - y\|^2$.*

1.2 Stochastic accelerated proximal gradient

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

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 $i = 1, \ldots, n$, 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 1.1 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.

3

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 $i = 1, \ldots, n$ it satisfies $0 = f_i(\bar{x})$.*

Consequently, each of the F_i satisfies Assumption 1.2 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 1.17) 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 1.15 (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) ...

This part is not finished yet.

4

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

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

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 $\frac{4 L_k}{L_{k-1}} \alpha_{k-1}^2 \geq 0$ 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:

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

5

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

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

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

$\square$

Remark 1.16. *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 1.13. 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 $k = 1, 2, \dots$.*

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 $L_0 \leq L_1 \leq \dots \leq L_k \leq \dots$ 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 $\tilde{\mu}_k = \rho \min_{i=1,\dots,k} L_{i-1}/L_i$ for some $\rho \in (0, 1)$.

{def:snapg-v2}

Definition 1.17 (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 $\{0, 1, 2, \cdots, n\}$. Initialize $v_{-1} = x_{-1}, \alpha_0 = 1$.* *Let $\tilde{\mu} \geq 0$ be a fixed constant.* *The SNAPG generates the sequence $(y_k, x_k, v_k)_{k\geq 0}$ such that for all $k \geq 0$ they satisfy:*

$$\alpha_k \in (0, 1) : (L_{k-1}/L_k)(1 - \alpha_k)\alpha_{k-1}^2 = \alpha_k\left(\alpha_k - \tilde{\mu}/L_k\right),$$

$$\tau_k = L_k(1 - \alpha_k)\left(L_k\alpha_k - \mu^{(I_k)}\right)^{-1},$$

$$y_k = (1 + \tau_k)^{-1}v_{k-1} + \tau_k(1 + \tau_k)^{-1}x_{k-1},$$

$$L_k > 0 : D_f(x_k, y_k) \leq L_k/2\|y_k - x_k\|^2,$$

$$x_k = T_{L_k}(y_k|F_{I_k}),$$

$$v_k = x_{k-1} + \alpha_k^{-1}(x_k - x_{k-1}).$$

6

Remark 1.18. $\tilde{\mu}_k, L_k$ *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 seems 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 auxilary 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 [?] for an implementations 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 1.19 (properties of the iterates)**.** *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 1.17. 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 1.17, it has

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

7

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 1.17 it has directly:

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

$\square$

Theorem 1.20 (SNAPG-V2 one step convergence)**.** *Let F satisfies assumption 1.14. Suppose that an algorithm satisfying Definition 1.17 uses this F. Let $\mathbb{E}_k$ denotes the expectation conditioned on $I_0, I_1, \ldots, I_{k-1}$. Then, for all $k \geq 1$, it has the following inequality*

$$
\mathbb{E}_k \left[F_{I_k}(x_k) \right] - F(\bar{x}) + \mathbb{E}_k \left[\frac{L_k \alpha_k^2}{2} \|\bar{x} - v_k\|^2 \right]
$$

$$
\leq (1 - \alpha_k) \left(\mathbb{E}_k \left[F_{I_k}(x_{k-1}) \right] - F(\bar{x}) + \mathbb{E}_k \left[\frac{\alpha_{k-1}^2 L_{k-1}}{2} \right] \|v_{k-1} - \bar{x}\|^2 \right)
$$

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

And for $k = 0$, it has

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

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. The proof is long so, we use letters and subscript under relations such as $=, \geq$ 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 1.7 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 1.9 with $z = z_k$ on F_i.

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

(d) Prove in Lemma 1.19 (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 1.19 (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 $k \geq 1$:

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

9

(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 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:

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

And we have the following chain of equalities:

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

10

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

$$+ \frac{(L_k \alpha_k - \mu^{(i)})\mu^{(i)}(1 - \alpha_k)}{(L_k - \mu^{(i)})} \langle \bar{x} - v_{k-1}, \bar{x} - x_{k-1} \rangle$$

$$\overset{=}{_{(g)}} \frac{(\alpha_k - 1)\mu^{(i)}\left(L_k \alpha_k - \mu^{(i)}\right)}{2\left(L_k - \mu^{(i)}\right)} \|\bar{x} - x_{k-1}\|^2$$

$$+ \left(\frac{\left(L_k \alpha_k - \mu^{(i)}\right)\mu^{(i)}(\alpha_k - 1)}{2(L_k - \mu^{(i)})} + \frac{\alpha_k(\tilde{\mu} - \mu^{(i)})}{2} \right) \|\bar{x} - v_{k-1}\|^2$$

$$+ \frac{\alpha_{k-1}^2 L_{k-1}(1 - \alpha_k)}{2} \|\bar{x} - v_{k-1}\|^2 + \frac{(L_k \alpha_k - \mu^{(i)})\mu^{(i)}(1 - \alpha_k)}{(L_k - \mu^{(i)})} \langle \bar{x} - v_{k-1}, \bar{x} - x_{k-1} \rangle$$

$$= \frac{(\alpha_k - 1)\mu^{(i)}\left(L_k \alpha_k - \mu^{(i)}\right)}{2\left(L_k - \mu^{(i)}\right)} \left(\|\bar{x} - x_{k-1}\|^2 + \|\bar{x} - v_{k-1}\|^2 - 2\langle \bar{x} - v_{k-1}, \bar{x} - x_{k-1} \rangle \right)$$

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

$$= \frac{(\alpha_k - 1)\mu^{(i)}\left(L_k \alpha_k - \mu^{(i)}\right)}{2\left(L_k - \mu^{(i)}\right)} \|x_{k-1} - v_{k-1}\|^2$$

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

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

$$0 \leq \alpha_k F_i(\bar{x}) + (1 - \alpha_k) F_i(x_{k-1}) - F_i(x_k)$$

$$- \frac{L_k}{2} \|z_k - x_k\|^2 + \frac{(\alpha_k - 1)\mu^{(i)}\left(L_k \alpha_k - \mu^{(i)}\right)}{2\left(L_k - \mu^{(i)}\right)} \|x_{k-1} - v_{k-1}\|^2$$

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

$$\overset{=}{_{(e)}} \alpha_k F_i(\bar{x}) + (1 - \alpha_k) F_i(x_{k-1}) - F_i(x_k)$$

$$- \frac{L_k \alpha_k^2}{2} \|\bar{x} - v_k\|^2 + \frac{(\alpha_k - 1)\mu^{(i)}\left(L_k \alpha_k - \mu^{(i)}\right)}{2\left(L_k - \mu^{(i)}\right)} \|x_{k-1} - v_{k-1}\|^2$$

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

$$= (\alpha_k - 1)F_i(\bar{x}) + (1 - \alpha_k) F_i(x_{k-1}) - F_i(x_k) + F_i(\bar{x})$$

$$- \frac{L_k \alpha_k^2}{2} \|\bar{x} - v_k\|^2 + \frac{(\alpha_k - 1)\mu^{(i)}\left(L_k \alpha_k - \mu^{(i)}\right)}{2\left(L_k - \mu^{(i)}\right)} \|x_{k-1} - v_{k-1}\|^2$$

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

11

$$= (1 - \alpha_k) \left(F_i(x_{k-1}) - F_i(\bar{x}) + \frac{\alpha_{k-1}^2 L_{k-1}}{2} \|v_{k-1} - \bar{x}\|^2 \right)$$

$$- \left(F_i(x_k) - F_i(\bar{x}) + \frac{L_k \alpha_k^2}{2} \|\bar{x} - v_k\|^2 \right)$$

$$+ \frac{\alpha_k(\tilde{\mu} - \mu^{(i)})}{2} \|\bar{x} - v_{k-1}\|^2 + \frac{(\alpha_k - 1)\mu^{(i)} \left(L_k \alpha_k - \mu^{(i)} \right)}{2 \left(L_k - \mu^{(i)} \right)} \|x_{k-1} - v_{k-1}\|^2.$$

Recall that $i = I_k$ is the random variable from Definition 1.17. Rearranging the last expression in the above equality chain can be conveniently written as

$$F_{I_k}(x_k) - F_{I_k}(\bar{x}) + \frac{L_k \alpha_k^2}{2} \|\bar{x} - v_k\|^2$$

$$\leq (1 - \alpha_k) \left(F_{I_k}(x_{k-1}) - F_{I_k}(\bar{x}) + \frac{\alpha_{k-1}^2 L_{k-1}}{2} \|v_{k-1} - \bar{x}\|^2 \right) \tag{1.2}$$

$$+ \frac{\alpha_k(\tilde{\mu} - \mu^{(I_k)})}{2} \|\bar{x} - v_{k-1}\|^2 + \frac{(\alpha_k - 1)\mu^{(I_k)} \left(L_k \alpha_k - \mu^{(I_k)} \right)}{2 \left(L_k - \mu^{(I_k)} \right)} \|x_{k-1} - v_{k-1}\|^2.$$

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:

$$\mathbb{E}_k \left[F_{I_k}(x_k) - F_{I_k}(\bar{x}) + \frac{L_k \alpha_k^2}{2} \|\bar{x} - v_k\|^2 \right]$$

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

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

$$\mathbb{E}_k \left[(1 - \alpha_k) \left(F_{I_k}(x_{k-1}) - F_{I_k}(\bar{x}) + \frac{\alpha_{k-1}^2 L_{k-1}}{2} \|v_{k-1} - \bar{x}\|^2 \right) \right]$$

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

$$\underset{\text{(1)}}{=} (1 - \alpha_k) \left(\mathbb{E}_k \left[F_{I_k}(x_{k-1}) \right] - \mathbb{E}_k[F_{I_k}(\bar{x})] + \mathbb{E}_k \left[\frac{\alpha_{k-1}^2 L_{k-1}}{2} \right] \|v_{k-1} - \bar{x}\|^2 \right)$$

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

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

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

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

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

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 1.13) 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:

$$\mathbb{E}_k \left[F_{I_k}(x_k) \right] - F(\bar{x}) + \mathbb{E}_k \left[\frac{L_k \alpha_k^2}{2} \|\bar{x} - v_k\|^2 \right]$$

$$\leq (1 - \alpha_k) \left(\mathbb{E}_k \left[F_{I_k}(x_{k-1}) \right] - F(\bar{x}) + \mathbb{E}_k \left[\frac{\alpha_{k-1}^2 L_{k-1}}{2} \right] \|v_{k-1} - \bar{x}\|^2 \right)$$

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

Finally, we show the base case. When $k = 0$, by assumption it had $\alpha_0 = 1$ hence τ_0 in Definition 1.17 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 1.7 with $z = z_0$ and, assume a successful line search with L_0, it yields:

$$0 \leq F_{I_0}(z_0) - F_{I_0}(x_0) - \frac{L_0}{2} \|z_0 - x_0\|^2 + \frac{L_0 - \mu^{(I_0)}}{2} \|z_0 - y_0\|^2$$

$$= F_{I_0}(\bar{x}) - F_{I_0}(x_0) - \frac{L_0}{2} \|\bar{x} - x_0\|^2 + \frac{L_0 - \mu^{(I_0)}}{2} \|\bar{x} - v_{-1}\|^2.$$

Re-arranging and taking the expectation it yields:

$$\mathbb{E} \left[F_{I_0}(x_0) - F_{I_0}(\bar{x}) + \frac{L_0}{2} \|\bar{x} - x_0\|^2 \right] \underset{(h)}{=} \mathbb{E} \left[F_{I_0} \right] - F(\bar{x}) + \frac{L_0}{2} \mathbb{E} \left[\|\bar{x} - x_0\|^2 \right]$$

$$\leq \frac{L_0 - \mathbb{E} \left[\mu^{(I_0)} \right]}{2} \|\bar{x} - v_{-1}\|^2$$

$$= \frac{L_0 - \mu}{2} \|\bar{x} - v_{-1}\|^2.$$

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

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

13

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

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

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

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

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

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

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

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

$$= \frac{\left(L_k\alpha_k - \mu^{(i)}\right)^2}{2(L_k - \mu^{(i)})} - \frac{L_k\alpha_k(\alpha_k - \mu^{(i)}/L_k)}{2} + \frac{L_k\alpha_k(\alpha_k - \mu^{(i)}/L_k)}{2} - \frac{L_k\alpha_k(\alpha_k - \tilde{\mu}/L_k)}{2}$$

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

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

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

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

$\square$

1.3 convergence rate of the algorithm under various circumstances

1.4 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 1.20 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] H. H. BAUSCHKE AND P. L. COMBETTES, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books in Mathematics, Springer International Publishing, Cham, 2017.