Lorum ipsum

Author 1 Name, Author 2 Name [*]

July 1, 2025

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

Abstract

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

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

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

1 Nesterov's Accelerated Gradient

1.1 In preparations

{ass:smooth-plus-nonsmooth}

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

{ass:smooth-plus-nonsmooth-x}

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

[*] Subject type, Some Department of Some University, Location of the University, Country. E-mail: `author.name@university.edu`.

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*

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

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
{def:rwapg-seq} Divergence has for all $x, y \in \mathbb{R}^n$: $\mu/2\|x - y\|^2 \leq D_f(x, y) \leq L/2\|x - y\|^2$.

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

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

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

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

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

{thm:pg-ineq} The following theorems are critical in analyzing the behavior of algorithm in Definition 1.9.

Theorem 1.10 (proximal gradient inequality) *Let function F satisfies Assumption 1.1, so it's $\mu \geq 0$ strongly convex. For all $x \in \mathbb{R}^n$, define $x^+ = T_L(x)$, then there exists* *a $B \ge 0$ such that $D_f(x^+, x) \le B/2\|x^+ - x\|^2$. Then, for all $z \in \mathbb{R}^n$ it satisfies proximal gradient inequality at point x:*

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

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

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

{thm:jesen}

Theorem 1.12 (Jensen's inequality) *Let $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a $\mu \ge 0$ 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*

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

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

1.2 A compact argument for the convergence of NAPG with proximal gradient

Here, the abbreviation "NAPG" stands for "Nesterov Acceleration Proximal Gradient". It's made in acknowledgement of Algorithm 2.2.36 in Nesterov's book [3] and its extension known as FISTA in the literature. The following theorem provides a complete proof for the convergence rate of algorithms implementing Definition 1.9, which is equivalent to NAG, or NAPG. It made use of Definition 1.7 which accommodates a relaxed sequence compared to the usual sequence that gives the optimal convergence rate.

{thm:onestep-napg-cnvg}

Theorem 1.14 (one step convergence claim of NAPG) *Let $F = f + g$ satisfies Assumption 1.1 for some $L > \mu \ge 0$. Let the sequence $(\alpha_k)_{k\ge0}$ be an R-WAPG sequence (Definition 1.7). Suppose that the iterates sequence $(x_k, y_k, v_k)_{k\ge0}$ satisfy NAPG in similar triangle form (Definition 1.9) with initial guesses $v_{-1}, x_{-1} \in \mathbb{R}^n$. Then for all $k \ge 1$, the* *following inequality is true for all $\bar{x} \in \mathbb{R}^n$:*

$$\begin{aligned}&- F(\bar{x}) + F(x_k) + \frac{L_k\alpha_k^2}{2}\|\bar{x} - v_k\|^2 \\ &\le \max\left(1, \frac{L_k\rho_{k-1}}{L_{k-1}}\right)(1-\alpha_k)\left(F(x_{k-1}) - F(\bar{x}) + \frac{\alpha_{k-1}^2 L_{k-1}}{2}\|\bar{x} - v_{k-1}\|^2\right).\end{aligned}$$

If in addition, we choose $\alpha_0 = 1$, and let $x_{-1} = v_{-1}$, then a base case of the inequality is:

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

Proof. The proof is very intense algebraically hence before we step into it, we present the following intermediate results in advance to their proofs given at the end.

For all $k \ge 0$, define $z_k = \alpha_k\bar{x} + (1-\alpha_k)x_{k-1}$.

(a) Theorem 1.10, with $z = z_k$, $k \ge 0$. We can use it because $F = f + g$ satisfies Assumption 1.1.

(b) Jensen's inequality (Theorem 1.12), with $z = z_k$, for $k \ge 0$. We can use it because F is $\mu \ge 0$ strongly convex.

(c) Definition 1.7 which has $\rho_k(1-\alpha_{k+1})\alpha_k^2 = \alpha_{k+1}(\alpha_{k+1} - \mu/L_k)$ for $k \ge 0$.

(d) The equality $z_k - y_k = (L_k - \mu)^{-1}((L_k\alpha_k - \mu)(\bar{x} - v_k) + \mu(1-\alpha_k)(\bar{x} - x_{k-1}))$ for all $k \ge 0$, it comes from Definition 1.9.

(e) The equality $z_k - x_k = \alpha_k(\bar{x} - v_k)$ for all $k \ge 0$ it comes from Definition 1.9.

(f) Using basic algebra, we have the following equality:

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

(g) Using (c), we have the following equality:

$$(\forall k \ge 1)\ \frac{1}{2}\left(\frac{(L_k\alpha_k - \mu)^2}{L_k - \mu} - \alpha_{k-1}^2\rho_{k-1}L_k(1-\alpha_k)\right) = \frac{\mu(L_k\alpha_k - \mu)(\alpha_k - 1)}{2(L_k - \mu)}.$$

(h) Definition 1.7 determine the inequality:

$$(\forall k \ge 1)\ \frac{\mu(L_k\alpha_k - \mu)(\alpha_k - 1)}{2(L_k - \mu)} \le 0.$$

With intermediate results (a) to (h), presented above, the proof of the theorem come smoothly from a chain of inequalities and equalities. The overall proof now follows. Start with the (a), the proximal gradient inequality it has:

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

Using the chain of equality below:

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

The inequality from previously simplifies, and it has:

$$0 \le \alpha_k F(\bar{x}) + (1-\alpha_k)F(x_{k-1}) - F(x_k) + \frac{\alpha_{k-1}^2 L_k \rho_{k-1}(1-\alpha_k)}{2}\|\bar{x} - v_{k-1}\|^2 - \frac{L_k}{2}\|z_k - x_k\|^2$$

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

Finally, for the base case, when $\alpha_0 = 1$, it has $y_0 = v_{-1} = x_{-1}$, and it makes $z_0 = \bar{x}$ therefore this makes the proximal gradient inequality into:

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

Going back to prove the intermediate results, the following will be useful. From Definition 1.9 it has for all $k \ge 0$

{thm:onestep-napg-cnvg-i}

$$
\tau_k = L_k(1-\alpha_k)(L_k\alpha_k - \mu)^{-1}. \tag{i}
$$

Then it has:

{thm:onestep-napg-cnvg-j}

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

And also

{thm:onestep-napg-cnvg-k}
$$\tau_k(1+\tau_k)^{-1} \underset{(i),(j)}{=} \frac{L_k(1-\alpha_k)}{L_k\alpha_k-\mu}\frac{L_k\alpha_k-\mu}{L_k-\mu} = \frac{L_k(1-\alpha_k)}{L_k-\mu}. \tag{k}$$

Proof of (d) For all $k \geq 1$, from Definition 1.9 it has

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

Recall the definition for z_k at the start of the proof and, use the above results it yields:

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

Proof of (e). The proof is direct using the equality with x_k in Definition 1.9.

$$\begin{aligned}
z_k - x_k &= \alpha_k\bar{x} + (1-\alpha_k)x_{k-1} - x_k \\
&= \alpha_k\bar{x} + x_{k1} - x_k - \alpha_kx_{k-1} \\
&= \alpha_k(\bar{x} - \alpha_k^{-1}(x_k - x_{k-1}) - x_{k-1}) \\
&= \alpha_k(\bar{x} - v_k).
\end{aligned}$$

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

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

Proof of (g) The proof is direct:

$$
\begin{aligned}
\frac{(L_k\alpha_k-\mu)^2}{2(L_k-\mu)} - \frac{\alpha_{k-1}^2 L_k\rho_{k-1}(1-\alpha_k)}{2} &\underset{\text{(c)}}{=} \frac{(L\alpha_k-\mu)^2}{2(L_k-\mu)} - \frac{L_k\alpha_k(\alpha_k-\mu/L_k)}{2} \\
&= \frac{1}{2(L_k-\mu)}\left((L_k\alpha_k-\mu)^2 - (L_k-\mu)L_k\alpha_k(\alpha_k-\mu/L_k)\right) \\
&= \frac{1}{2(L_k-\mu)}\left((L_k\alpha_k-\mu)^2 - (L_k-\mu)\alpha_k(L_k\alpha_k-\mu)\right) \\
&= \frac{L_k\alpha_k-\mu}{2(L_k-\mu)}\left(L_k\alpha_k-\mu-(L-\mu)\alpha_k\right) \\
&= \frac{L_k\alpha_k-\mu}{2(L_k-\mu)}\left(\mu\alpha_k-\mu\right) \\
&= \frac{(L\alpha_k-\mu)\mu(\alpha_k-1)}{2(L_k-\mu)}.
\end{aligned}
$$

Proof of (h). For all $k \geq 1$, by (c), the definition of the R-WAPG sequence, $\alpha_k \in (\mu/L_k, 1)$, then it has $L_k\alpha_k \in (\mu, L_k)$, so $L_k\alpha_k - \mu > 0$, and $\alpha_k - 1 < 0$. Finally, we have $L_k \geq \mu$, therefore, the fraction is negative. ■

1.3 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 1.15 (sum of many) Define $F := g + (1/n)\sum_{i=1}^{n} f_i$, assume that $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ are all $K^{(i)}$ smooth and $\mu^{(i)} \geq 0$ strongly convex function such that $K^{(i)} > \mu^{(i)}$ and, g :

$\mathbb{R}^n \to \overline{\mathbb{R}}$ is a closed convex proper function. Consequently, the function f cane be written as $F = g + f$ where $f = (1/n)\sum_{i=1}^{n} f_i$ and, it satisfies Assumption 1.1 with $L = \max_{i=1,\ldots,n} K^{(i)}$ and $\mu = (1/n)\sum_{i=1}^{n} \mu^{(i)}$.

This assumption is stronger than Assumption 1.1. The interpolation hypothesis from Machine Learning stated that the model has the capacity to perfect fit all the observed data. {ass:interp-hypothesis}

Assumption 1.16 (interpolation hypothesis) Suppose that $F := f + (1/n)\sum_{i=1}^{n} f_i$ satisfying Assumption 1.15. In addition, assuming that it has $0 = \inf_x F(x)$ and, there exists some $\bar{x} \in \mathbb{R}^n$ such that for all $i = 1, \ldots, n$ it satisfies $0 = f_i(\bar{x})$. Obviously, all such $\bar{x}$ forms the set of minimizers of F.

{def:snapg-v1} **Definition 1.17 (SNAPG-V1 DOESN'T WORK WELL)**
Let F satisfies Assumption 1.15. Let $(I_k)_{k\ge 0}$ be a list of i.i.d random variables uniformly sampled from set $\{0, 1, 2, \cdots, n\}$. Initialize $v_{-1} = x_{-1}, \alpha_0 = 1$. The SNAPG generates the sequence $(y_k, x_k, v_k)_{k\ge 0}$ such that for all $k \ge 0$ they satisfy:

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

Remark 1.18 The sequence α_k is a random variable because it changes according to variable I_k at the current iteration. This decision is made because if Theorem 1.10 needs to hold during each iteration of the algorithm.

Theorem 1.19 (SNAPG one step inequality INVALIDATED) *Let sequences of iterates $(y_k, x_k, v_k)_{k\ge 0}$ satisfy SNAPG (Definition 1.17). Denote $\mathbb{E}_k[\cdot]$ to be the conditional expectation on $I_0, I_1, \ldots, I_{k-1}$. Then, for all $k \ge 1$ the following inequality is true:*

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

Proof. For one step, a specific function F_{I_k} is sampled. F_{I_k} is $K^{(I_k)} > \mu^{(I_k)} \ge 0$ strongly convex and smooth. The iterates x_k, v_k is a function of random variable I_k, conditioned on all $I_{k-1}, I_{k-2}, \ldots, I_0$ The sequence α_k is a sequence of random variable and, the relations between α_k, α_{k-1} is given by:

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

The relationship $(L_{k-1}/L_k)(1-\alpha_k)\alpha_{k-1}^2 = \alpha_k(\alpha_k - \mu^{(I_k)}/L_k)$ is an instance of Definition 1.7 where $\rho_{k-1} = L_{k-1}/L_k$. Therefore, the one-step convergence result of 1.14 applies to function F_{I_k} with sequence α_k and, it yields:

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

We take the conditional expectation on the LHS and, it has

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

At label (a), we make use of interpolation hypothesis which makes $\mathbb{E}_k[F_{I_k}(\bar{x})] = F(\bar{x})$. Next, taking the conditional expectation on the RHS it has

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

At (b), we need to make use of covariance between the two random variable... Consider random variables $X_k = (1-\alpha_k)$ and $Y_k = F_{I_k}(x_{k-1}) - F_{I_k}(\bar{x}) + \alpha_{k-1}L_{k-1}/2\|\bar{x} - v_{k-1}\|^2$ which are functions of I_k conditioned on α_{k-1} and, x_{k-1}, v_{k-1}. The conditional expectation of their product has $\mathbb{E}_k[X_kY_k] = \mathbb{E}_k[X_k]\mathbb{E}_k[Y_k] + \operatorname{Cov}[X_k, Y_k]$. Hence, taking the expectation doesn't follow.

■

{def:snapg-v2}

Definition 1.20 (SNAPG-V2) *Let F satisfies Assumption 1.15. Let $(I_k)_{k\geq 0}$ be a list of i.i.d random variables uniformly sampled from set $\{0, 1, 2, \cdots, n\}$. Initialize $v_{-1} = x_{-1}, \alpha_0 = 1$. The SNAPG generates the sequence $(y_k, x_k, v_k)_{k\geq 0}$ such that for all $k \geq 0$ they satisfy:*

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

Theorem 1.21 (SNAPG-V2 one step convergence)

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. From Definiton 1.20, it has

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

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

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

■

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.