Linear Convergence of Stochastic Nesterov's Accelerated Proximal Gradient method under Interpolation Hypothesis, Truth or Just a Dream?

Author *

July 16, 2025

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

Abstract

This file is for communication purposes between collaborators. In brief, we think that the conditions required for linear convergence rate of Stochastic Nesterov's accelerated gradient (or proximal gradient) is too precarious to hold, even with interpolation hypothesis. Instead of attacking the problem head on, this file will characterize the conditions required for linear convergence rate. We place specific constraints on the random variable representing the error made when estimating gradient via some random variables.

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

1 Introduction

[1] Previously we got some results, but unfortunately it was incorrect and, it was impossible to recover from the mistake.

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

Does stochastic accelerated Nesterov's acceleration (SNAG) produces accelerated convergence rate (or, any type of convergence) when the Interpolation Hypothesis is true? **I don't think that it's true after some mistakes from previous version of the notes and careful investigations.** In this file we develop some sufficient conditions for Linear convergence of (SNAG). We will give explanations on why we don't think this is necessarily true.

When we use stochastic gradient to approximate the true graduate, it has an error. Fix some $x \in \mathbb{R}^n$, let $\tilde{\nabla} f(x)$ be an estimate of $\nabla f(x)$, the error we consider is $\mathbb{E}\|\nabla f(x) - \nabla f(x)\|$. To make the algebra simpler, we assume that the algorithm produced the next iterates $\tilde{x}$ by a step of gradient descent, and the error of the expectation satisfies a relative error conditions of the form

$$\frac{\mathbb{E}\left[\left\|\nabla f(x) - \tilde{\nabla} f(x)\right\| \|z - \tilde{x}\|\right]}{\mathbb{E}\left[\|x - \tilde{x}\|\|z - \tilde{x}\|\right]} = \epsilon.$$

Where the variable z will be explained later. We will show that, the value of ϵ must decreases at a rate convergence relative to the Nesterov's accelerated sequence, under the standard Framework of analysis similar to what is in the literature. Take note that usually in the literature, people analyze the quantity $\mathbb{E}\left\|\tilde{\nabla} f(x) - \tilde{\nabla} f(y)\right\|^2$ for stochastic gradient type of method. The above expression is drastically different from what we usually have in the literature.

2 In preparations

Unless specifically specified in the context, we use the following notations. Π_C denotes the projection onto a set C. Let $A \in \mathbb{R}^{m\times n}$ be a matrix. $\sigma_{\min}(A)$ denotes the smallest non-zero absolute value of all singular values of A. Let $\|A\|$ denotes the spectral norm of the matrix A. I denotes the identity operator.

When two expressions are connected via non-trivial results, it's expressed with $\underset{(\cdot)}{=}, \underset{(\cdot)}{\geq}$ where $(\cdot)$ is a label of some intermediate results immediately before it, or explained right after a chain of expressions. If the label is letter, like: (a), (b), ..., then they are stated in advanced at the start of the proof and they are usually non-trivial results. These labels are reused in every proof. If the label is circled numbers, like: ①, ②,... they are explained right after the chain of relations, and they are often reused right after their explanations.

{def:pg-opt}

2.1 Basic definitions

Definition 2.1 (Proximal gradient operator)**.** *Suppose* $F = f + g$ *with* $\operatorname{ri}(\operatorname{dom} f) \cap \operatorname{ri}(\operatorname{dom} g) \neq \emptyset$*, and* f *is a differentiable function. Let* $\beta > 0$*. Then, we define the proximal gradient operator* T_β *as*

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

Remark 2.2. *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 2.3 (Bregman Divergence)**.** *Let* $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ *be a differentiable function. Then, for all the Bregman divergence* $D_f : \mathbb{R}^n \times \operatorname{dom} \nabla f \to \mathbb{R}$ *is defined as:*

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

Remark 2.4. *If,* f *is* $\mu \geq 0$ *strongly convex and* L *Lipschitz smooth then, its Bregman Divergence has for all* $x, y \in \mathbb{R}^n$*:* $\mu/2\|x - y\|^2 \leq D_f(x, y) \leq L/2\|x - y\|^2$*. We note that usually the Bregman Divergence is used with a Legendre function, but in here, we do not assume that* f *has to be Legendre.*

{def:lip-smooth-and-scnvx}

Definition 2.5 (Lipschitz smoothness and strongly convex)**.** *A differentiable function* $f : \mathbb{R}^n \to \mathbb{R}$ *is* L *lipschitz smooth and,* μ *strong convex for some* $L > \mu \geq 0$ *if and only if for all* $x, y \in \mathbb{R}^n$ *it satisfies the inequality*

$$\frac{\mu}{2}\|x - y\|^2 \leq D_f(x, y) \leq \frac{L}{2}\|x - y\|^2.$$

2.2 Important inequalities

{ass:smooth-plus-nonsmooth}

Assumption 2.6. *Suppose that* $F = f + g$ *where* f, g *are both convex, proper and closed. In addition, assume* f *is* $L > \mu \geq 0$ *Lipschitz smooth and strongly convex satisfying Definition 2.5.*

{thm:jesen}

Theorem 2.7 (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 2.8. *If* x, y *is out of* $\operatorname{dom} F$*, the inequality still work by convexity.*

{lemma:inex-pg-ineq-proto}

Lemma 2.9 (inexact proximal gradient inequality prototype)**.** *Let $F = f + g$ satisfies Assumption 2.6. Fix some $x \in \mathbb{R}^n$, and suppose that an error: w is made when estimating the proximal to obtain $\tilde{x}$ at x such that it's characterized by*

$$w \in \partial \left[z \mapsto g(z) + \langle \nabla f(x), z - x \rangle + \frac{B}{2}\|z - x\|^2 \right] (\tilde{x})$$

And in addition, assume that there exists some $B \geq 0$ such that $D_f(\tilde{x}, x) \leq \frac{B}{2}\|\tilde{x} - x\|^2$. Then, for all $z \in \mathbb{R}^n$ it satisfies:

$$\frac{B}{2}\|z - \tilde{x}\|^2 \leq F(z) - F(\tilde{x}) + \frac{\max(B, L) - \mu}{2}\|z - x\|^2 - \langle w, z - \tilde{x} \rangle.$$

Proof. The proof is direct algebra. Let $h = z \mapsto g(z) + \langle \nabla f(x), z - x \rangle + B/2\|z - x\|^2$. h is a B strongly convex function, using the subgradient inequality of a strongly convex function it has for all $z \in \mathbb{R}^n$:

$$\begin{aligned}
\frac{B}{2}\|z - \tilde{x}\|^2 &\leq h(z) - h(\tilde{x}) - \langle w, z - \tilde{x} \rangle \\
&= \left(g(z) + \langle \nabla f(x), z - x \rangle + \frac{B}{2}\|z - x\|^2 \right) \\
&\quad - \left(g(\tilde{x}) + \langle \nabla f(x), \tilde{x} - x \rangle + \frac{B}{2}\|\tilde{x} - x\|^2 \right) - \langle w, z - \tilde{x} \rangle \\
&= \left(g(z) + f(z) - f(z) + \langle \nabla f(x), z - x \rangle + \frac{B}{2}\|z - x\|^2 \right) \\
&\quad - \left(g(\tilde{x}) + f(\tilde{x}) - f(\tilde{x}) + \langle \nabla f(x), \tilde{x} - x \rangle + \frac{B}{2}\|\tilde{x} - x\|^2 \right) - \langle w, z - \tilde{x} \rangle \\
&= \left(F(z) - D_f(z, x) + \frac{B}{2}\|z - x\|^2 \right) \\
&\quad - \left(F(\tilde{x}) - D_f(\tilde{x}, x) + \frac{B}{2}\|\tilde{x} - x\|^2 \right) - \langle w, z - \tilde{x} \rangle \\
&\underset{①}{\leq} F(z) - F(\tilde{x}) + \frac{\max(B, L) - \mu}{2}\|z - x\|^2 - 0 - \langle w, z - \tilde{x} \rangle
\end{aligned}$$

At ①, we used the fact that f is $L > \mu \geq 0$ Lipschitz smooth and strongly convex therefore it has for all $y \in \mathbb{R}^n$:

$$0 \leq \frac{L}{2}\|z - y\|^2 - D_f(z, y) \leq \frac{L - \mu}{2}\|z - y\|^2.$$

□

{lemma:inex-pg-ineq}

Lemma 2.10 (inexact proximal gradient inequality)**.** *Let $F = f + g$ satisfies Definition 2.5 with $L > \mu \geq 0$. Let $x \in \mathbb{R}^n$ be fixed. Suppose an inexact evaluation of proximal gradient operator at x yield an approximation $\tilde{x}$ such that:*

$$(\exists w)(\exists \epsilon) : w \in \partial \left[z \mapsto g(z) + \langle \nabla f(x), z - x \rangle + \frac{B}{2}\|z - x\|^2 \right] (\tilde{x}), \|w\| \leq \epsilon \|x - \tilde{x}\|.$$

Suppose that there exists some $B \geq 0$ such that $D_f(\tilde{x}, x) \leq \frac{B}{2}\|\tilde{x} - x\|^2$. Then, for all $z \in \mathbb{R}^n$ it satisfies the inequality:

$$0 \leq F(z) - F(\tilde{x}) + \frac{\max(B, L) - \mu}{2}\|z - x\|^2 - \frac{B - \epsilon}{2}\|z - \tilde{x}\|^2 + \frac{\epsilon}{2}\|x - \tilde{x}\|^2.$$

Proof. The error w satisfies Lemma 2.9 hence, it has for all $z \in \mathbb{R}^n$ the inequality:

$$\begin{aligned}\frac{B}{2}\|z - \tilde{x}\|^2 &\leq F(z) - F(\tilde{x}) + \frac{\max(B, L) - \mu}{2}\|z - x\|^2 - 0 - \langle w, z - \tilde{x} \rangle \\ &\underset{①}{\leq} F(z) - F(\tilde{x}) + \frac{\max(B, L) - \mu}{2}\|z - x\|^2 + \epsilon\|x - \tilde{x}\|\|z - \tilde{x}\|.\end{aligned}$$

At ①, we used Cauchy inequality. Continuing it has

$$\begin{aligned}0 &\leq F(z) - F(\tilde{x}) + \frac{\max(B, L) - \mu}{2}\|z - x\|^2 + \epsilon\|x - \tilde{x}\|\|z - \tilde{x}\| - \frac{B}{2}\|z - \tilde{x}\|^2 \\ &\quad - \frac{\epsilon}{2}\|x - \tilde{x}\|^2 - \frac{\epsilon}{2}\|z - \tilde{x}\|^2 + \frac{\epsilon}{2}\|x - \tilde{x}\|^2 + \frac{\epsilon}{2}\|z - \tilde{x}\|^2 \\ &= F(z) - F(\tilde{x}) + \frac{\max(B, L) - \mu}{2}\|z - x\|^2 - \frac{1}{2}\left(\sqrt{\epsilon}\|z - \tilde{x}\| - \sqrt{\epsilon}\|x - \tilde{x}\|\right)^2 - \frac{B}{2}\|z - \tilde{x}\|^2 \\ &\quad + \frac{\epsilon}{2}\|x - \tilde{x}\|^2 + \frac{\epsilon}{2}\|z - \tilde{x}\|^2 \\ &= F(z) - F(\tilde{x}) + \frac{\max(B, L) - \mu}{2}\|z - x\|^2 - \frac{B - \epsilon}{2}\|z - \tilde{x}\|^2 + \frac{\epsilon}{2}\|x - \tilde{x}\|^2.\end{aligned}$$

□

Remark 2.11. *Usually in practice, the precise value of $F(\tilde{x})$ is never known and, function value is also a random variable, therefore, B cannot be easily determined via $D_f(\tilde{x}, x)$. In that case we can only choose some $B \geq L$ which gives:*

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

The smallest possible choice for ϵ when $x \neq \tilde{x}$ is $\epsilon = \|w\|/\|x - \tilde{x}\|$ and, if $x = \tilde{x}$ then $\epsilon = 0$ is the smallest.

Note, an inexact evaluation of the proximal gradient operator can be caused by an inexact gradient on the smooth part. Suppose that one take $\tilde{\nabla} f(x)$ to be an estimate of $\nabla f(x)$ and use it for the proximal gradient operator to produce $\tilde{x}$, then:

$$\begin{aligned} \mathbf{0} &\in \partial g(\tilde{x}) + \tilde{\nabla} f(x) + B(\tilde{x} - x) && (2.1)\\ &= \partial g(\tilde{x}) + \tilde{\nabla} f(x) - \nabla f(x) + \nabla f(x) + B(\tilde{x} - x) && (2.2)\\ \{\text{eqn:stoch-grad-err-vec}\} \quad \iff \nabla f(x) - \tilde{\nabla} f(x) &\in \partial g(\tilde{x}) + \nabla f(x) + B(\tilde{x} - x). && (2.3) \end{aligned}$$

In this case, it adds the interpretation that $w = \nabla f(x) - \tilde{\nabla} f(x)$. It fully characterizes the error made to estimate the true gradient $\nabla f(x)$. In that case, we have the equation:

$$\left\| \nabla f(x) - \tilde{\nabla} f(x) \right\| \|x - \tilde{x}\| = \epsilon \|x - \tilde{x}\|^2.$$

It's very unclear what LHS really is without additional details and assumptions. **We very much would like ϵ to be a constant to make the algebra possible when deriving the convergence rate of the algorithm.**

{lemma:stoch-pg-ineq} The following lemma gives a proximal gradient inequality when $\tilde{\nabla} f(x)$ is an estimate by some random variable, **and it is the precursor.**

Lemma 2.12 (stochastic proximal gradient inequality)**.** *Let $F = f + g$ satisfies Assumption 2.6. Fix any $x, z \in \mathbb{R}^n$. Suppose that, $\tilde{\nabla} f(x)$ is a random variable which estimates $\nabla f(x)$, cause proximal gradient operator to produce error w, and estimate $\tilde{x}$ as given by 2.9. Assume there exists $B \geq 0$ be constant such that $D_f(\tilde{x}, x) \leq B/2 \|x - \tilde{x}\|^2$. Then, it would have $w = \nabla f(x) - \tilde{\nabla} f(x)$. If in addition, there exists an $\epsilon \geq 0$:*

$$\mathbb{E}\left[\|w\| \|z - \tilde{x}\|\right] \leq \epsilon \mathbb{E}\left[\|x - \tilde{x}\| \|z - \tilde{x}\|\right].$$

Then it has:

$$0 \leq F(z) + \mathbb{E}F(\tilde{x}) + \frac{\max(B, L) - \mu}{2} \|z - x\|^2 - \frac{B - \epsilon}{2} \mathbb{E}\left[\|z - \tilde{x}\|^2\right] + \frac{\epsilon}{2} \mathbb{E}\left[\|x - \tilde{x}\|^2\right].$$

Proof. The reason for $w = \nabla f(x) - \tilde{\nabla} f(x)$ is explained in (2.3). Using Lemma 2.9, for any fixed z it has:

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

Take note that, since $w = \nabla f(x) - \tilde{\nabla} f(x)$ is a random variable, it determines that $\tilde{x}$ is also a random variable related to w. Here, x, z is not a random variable. We take the expectation

on both sides and move things all to the RHS then it has

$$\begin{aligned}
0 &\leq F(z) - \mathbb{E}F(\tilde{x}) + \frac{\max(B,L)-\mu}{2}\|z-x\|^2 + \mathbb{E}\left[\|w\|\|z-\tilde{x}\|\right] - \frac{B}{2}\mathbb{E}\|z-\tilde{x}\|^2 \\
&\leq F(z) - \mathbb{E}F(\tilde{x}) + \frac{\max(B,L)-\mu}{2}\|z-x\|^2 + \epsilon\mathbb{E}\left[\|x-\tilde{x}\|\|z-\tilde{x}\|\right] - \frac{B}{2}\mathbb{E}\|z-\tilde{x}\|^2 \\
&= F(z) - \mathbb{E}F(\tilde{x}) + \frac{\max(B,L)-\mu}{2}\|z-x\|^2 + \mathbb{E}\left[\epsilon\|w\|\|z-\tilde{x}\| - \frac{B}{2}\|z-\tilde{x}\|^2\right] \\
&= F(z) - \mathbb{E}F(\tilde{x}) + \frac{\max(B,L)-\mu}{2}\|z-x\|^2 \\
&\quad + \mathbb{E}\left[-\frac{1}{2}\left(\sqrt{\epsilon}\|x-\tilde{x}\| - \sqrt{\epsilon}\|z-\tilde{x}\|\right)^2 + \frac{\epsilon}{2}\|x-\tilde{x}\|^2 + \frac{\epsilon}{2}\|z-\tilde{x}\|^2 - \frac{B}{2}\|z-\tilde{x}\|^2\right] \\
&\leq F(z) - \mathbb{E}F(\tilde{x}) + \frac{\max(B,L)-\mu}{2}\|z-x\|^2 + \mathbb{E}\left[\frac{\epsilon}{2}\|x-\tilde{x}\|^2 - \frac{B-\epsilon}{2}\|z-\tilde{x}\|^2\right].
\end{aligned}$$

□

Remark 2.13. *In practice, is chosen in prior to satisfies $B \geq L$. In here, z, x is not a random variable, ϵ just a constant, but it's determined by z and x. Assuming $z \neq \tilde{x}$ and, $\tilde{x} \neq x$, then one of the smallest choice for it in this lemma is*

$$\frac{\mathbb{E}\left[\left\|\nabla f(x) - \tilde{\nabla} f(x)\right\| \|z-\tilde{x}\|\right]}{\mathbb{E}\left[\|x-\tilde{x}\|\|z-\tilde{x}\|\right]} = \epsilon.$$

Otherwise, $\epsilon = 0$ is an obvious choice and, it's the smallest. It depends on both z and x.

Of course, look, if there is no random variable and $\tilde{\nabla} f(x)$ is simply not a probabilistic estimate then the expectation is gone and $x \neq \tilde{x}$, it has:

$$\epsilon = \frac{\left\|\nabla f(x) - \tilde{\nabla} f(x)\right\|}{\|x-\tilde{x}\|}.$$

And in this case, it doesn't on z.

3 Stochastic/Inexact accelerated proximal gradient algorithm

The following defines the inexact proximal gradient operator where the gradient of the smooth part of the function is estimated. All algorithms satisfying the following definition will be referred to as Stochastic Nesterov's Accelerated Gradient (SNAG).

Definition 3.1 (inexact proximal gradient operator with relative error)**.** *Let $F = f + g$ satisfies Assumption 2.6, let $x \in \mathbb{R}^n$ be fixed. Suppose that $\tilde{\nabla} f(x)$ estimates $\nabla f(x)$. We define the inexact proximal gradient operator by the relationships between:*

(i) $\tilde{x} = \mathbf{T}_B^{(\epsilon)}(x|F)$ *is an inexact output of proximal gradient operator by evaluating on* $\tilde{\nabla} f(x)$.
(ii) $B \geq 0$ *is any constant such that it satisfies* $D_f(\tilde{x}, x) \leq B/2\|\tilde{x} - x\|^2$.
(iii) $\epsilon \geq 0$ *is a constant that quantifies the error of inexact evaluation.*

Then, we define the error ϵ by:

$$\epsilon = \begin{cases} \frac{\left\|\tilde{\nabla} f(x) - \nabla f(x)\right\|}{\|x - \tilde{x}\|} & \text{if } x \neq \tilde{x}, \\ 0 & \text{else } x = \tilde{x}. \end{cases}$$

And the inexact output is defined as:

$$\tilde{x} = \mathbf{T}_B(x|F) = \operatorname*{argmin}_{z \in \mathbb{R}^n} \left\{ g(z) + \left\langle \tilde{\nabla} f(x), z - x \right\rangle + \frac{B}{2}\|z - x\|^2 \right\}.$$

The inexact evaluation can be caused by a random variable. The definition that follows characterize algorithm in which the errors are related to a random variable that estimates the gradient of the objective function.

{def:stoch-pg-opt-rel-err}

Definition 3.2 (stochastic proximal gradient operator with relative error)**.** *Let $F = f + g$ satisfies Assumption 2.6. Let $x \in \mathbb{R}^n$ be fixed. Suppose that $\tilde{\nabla} f(x)$ is random variable and, it estimates $\nabla f(x)$. Then stochastic proximal gradient operator are the relationships between*

(i) $\tilde{x} = \widetilde{\mathbf{T}}_B(x|F)$, *an inexact output of proximal gradient operator by evaluating on* $\tilde{\nabla} f(x)$.
(ii) ϵ *an error which is pre-determined constant.*
(iii) *Any* $B \geq 0$ *such that it satisfies* $D_f(\tilde{x}, x) \leq B/2\|x - \tilde{x}\|^2$.

Here, ϵ satisfies:

$$\epsilon = \begin{cases} \frac{\mathbb{E}\left[\left\|\nabla f(x) - \tilde{\nabla} f(x)\right\| \|z - \tilde{x}\|\right]}{\mathbb{E}[\|x - \tilde{x}\| \|z - \tilde{x}\|]} & \text{if } x \neq \tilde{x} \ \wedge \ z \neq \tilde{x} \\ 0 & \text{else} \end{cases}$$

Then the inexact proximal gradient operator with relative error ϵ is the random variable defined as:

$$\tilde{x} = \widetilde{\mathbf{T}}_B(x|F) = \operatorname*{argmin}_{z \in \mathbb{R}^n} \left\{ g(z) + \left\langle \tilde{\nabla} f(x), z - x \right\rangle + \frac{B}{2}\|z - x\|^2 \right\}.$$

Remark 3.3. *For this definition of the stochastic proximal gradient operator, Lemma 2.12 is applicable.*

{def:SNAG}

Definition 3.4 (inexact/stochastic SNAG)**.** *Suppose that $F = f + g$ satisfies Assumption 2.6. Let $(\alpha_k)_{k\geq 0}$ be a sequence such that $\alpha_k \in (0, 1]$. Let $(\epsilon_k)_{k\geq 0}$ be a sequence of errors. Given initial conditions v_{-1}, x_{-1}. An algorithm satisfying the SNAG definition if it generates a sequence $(y_k, x_k, v_k)_{k\geq 0}$ if for all $k \geq 0$, the following conditions are satisfied:*

$$\begin{aligned}
\tau_k &= L(1-\alpha_k)\left(L\alpha_k - \mu\right)^{-1}, \\
y_k &= (1+\tau_k)^{-1} v_{k-1} + \tau_k (1+\tau_k)^{-1} x_{k-1}, \\
x_k &= \mathbf{T}_L(y_k|F) \text{ or }, \widetilde{\mathbf{T}}_L(y_k|F) \\
v_k &= x_{k-1} + \alpha_k^{-1}(x_k - x_{k-1}).
\end{aligned}$$

{def:relax-momen-seq}

The following definition gives a momentum sequence where it makes the derivation of the convergence rate easier.

Definition 3.5 (relaxed momentum sequence)**.** *Let $(\alpha_k)_{k\geq 0}$ be non-negative sequence. Let L, μ be some constant such that $L > \mu \geq 0$. It is a relaxed momentum sequence if the following conditions are satisfied:*

(i) *$\alpha_0 \in (0, 1]$ and for all $k \geq 1$, it satisfies that $\alpha_k \in (\mu/L, 1)$.*

{lemma:seq-properties}

Remark 3.6. *In the context of its usage, the constants L, μ are the Lipschitz smoothness constant and, strong convexity constant associated with a smooth and strongly convex function.*

Lemma 3.7 (properties of the sequence)**.**

3.1 building up the convergence results

In this section we derive a generic convergence results.

The following lemma states two important relationships on the iterates generated by Definition 3.4. Take note that it's only related to the iterates generated: x_k, y_k, v_k, it involves the sequence $(\alpha_k)_{k\geq 0}$, but the sequence can be anything in between $(0, 1]$ and these relations won't change.

{lemma:snag-identities}

Lemma 3.8 (SNAG identities)**.** *The iterates $(y_k, x_k, v_k)_{k\geq 0}$ satisfying Definition 3.4 satisfies for all $k \geq 1$ the identities:*

(i) $z_k - y_k = (L-\mu)^{-1}((L\alpha_k - \mu)(\bar{x} - v_{k-1}) + \mu(1-\alpha_k)(\bar{x} - x_{k-1}))$.
(ii) $z_k - x_k = \alpha_k(\bar{x} - v_k)$.

Proof. We prove (i) first. Recall the definitions of τ_k from Definition 3.4, it has:

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

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

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

Therefore, we establish the equality

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

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

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

To see item (ii), the proof is direct algebra:

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

□

{thm:snag-descent}

The following theorem given an inequality characterizing a descent relation for the SNAG algorithm.

Theorem 3.9 (SNAG descent lemma)**.** *Suppose the iterates sequence* $(y_k, x_k, v_k)_{k\geq 0}$ *are generated by algorithms satisfying Definition 3.4. Assume that*

(i) *it uses stochastic proximal gradient operator as in Definition 3.2,*
(ii) *it is initialized with* $v_{-1} = x_{-1}$, $\alpha_0 = 1$ *and, the momentum sequence* $(\alpha_k)_{k\geq 0}$ *satisfies Definition 3.5.*

Define $\mathbb{E}_k$ *to be the expectation conditioned on* $\tilde{\nabla} f(y_i)$ *for* $i = 1, 2, \ldots, k-1$*. Then for all* $k \geq 1$*, for all* $\bar{x} \in \mathbb{R}^n$ *it satisfies the inequality:*

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

Define $z_k = \alpha_k\bar{x} + (1-\alpha_k)x_{k-1}$*. Then, the error sequence* $(\epsilon_k)_{k\geq 0}$ *is given by:*

$$(\forall k \geq 1)\ \epsilon_k = \frac{\mathbb{E}_k\left[\left\|\tilde{\nabla} f(y_k) - \nabla f(y_k)\right\|\|\bar{x} - v_k\|\right]}{\mathbb{E}_k\left[\|\bar{x} - v_k\|\|z_k - y_k\|\right]}.$$

Proof. The following intermediate results will clear out some algebras, they are all proved by the end of the proof.

(a) For all $k \geq 1$, it has $\frac{\mu^2(1-\alpha_k)^2}{2(L-\mu)} - \frac{\mu\alpha_k(1-\alpha_k)}{2} = \frac{(\alpha_k-1)\mu(L\alpha_k-\mu)}{2(L-\mu)}$ using some algebra.

(b) We assumed that the sequence $(\alpha_k)_{k\geq 0}$ satisfies for all $k \geq 1$: $\rho_{k-1}(1-\alpha_k)\alpha_{k-1}^2 = \alpha_k(\alpha_k - \mu/L)$.

(c) Using (b) and some algebra, we have for all $k \geq 1$ the identity: $\frac{(L\alpha_k-\mu)^2}{2(L-\mu)} - \frac{\alpha_{k-1}^2 L\rho_{k-1}(1-\alpha_k)}{2} = \frac{(L\alpha_k-\mu)\mu(\alpha_k-1)}{2(L-\mu)}$.

(d) Using (a), and (c), we can derive for all $k \geq 1$, we have the following identity:

$$\begin{aligned}&-\frac{\mu\alpha_k(1-\alpha_k)}{2}\|\bar{x} - x_{k-1}\|^2 + \frac{L-\mu}{2}\|z_k - y_k\|^2\\ &= \frac{\alpha_{k-1}^2 L\rho_{k-1}(1-\alpha_k)}{2}\|\bar{x} - v_{k-1}\|^2 + \frac{(\alpha_k-1)\mu(L\alpha_k-\mu)}{2(L-\mu)}\|x_{k-1} - v_{k-1}\|^2.\end{aligned}$$

Using intermediate results (a), (b), (c), (d), we can prove the claim in just a few steps. For any fixed $\bar{x} \in \mathbb{R}^n$. Define $z_k = \alpha_k \bar{x} + (1-\alpha_k)x_{k-1}$ for all $k \geq 0$. Consider the case for all $k \geq 1$. Recall $\mathbb{E}_k$ is the expectation conditioned on all $\tilde{\nabla} f(y_i)$ for $i = 0, 1, \ldots, k-1$. We note that under this conditioning the only random variable is $\tilde{\nabla} f(y_k)$, so iterates x_{k-1}, v_{k-1}, y_k are not random variables, but x_k, and v_k are. The sequence $(\alpha_k)_{k\geq 0}, (\epsilon_k)_{k\geq 0}$ are also not random variables.

We use Lemma 2.12 with $x = y_k, z = z_k, \tilde{x} = x_k$ and, $B = L$ then it means:

$$
\begin{aligned}
0 &\leq F(z_k) - \mathbb{E}_k F(x_k) + \frac{L-\mu}{2}\|z_k - y_k\|^2 + \frac{\epsilon_k}{2}\mathbb{E}_k\left[\|y_k - x_k\|^2\right] \\
&\quad - \frac{L-\epsilon_k}{2}\mathbb{E}_k\left[\|z_k - x_k\|^2\right] \\
&\underset{①}{\leq} \alpha_k F(\bar{x}) + (1-\alpha_k)F(x_{k-1}) - \mathbb{E}_k F(x_k) - \frac{\mu\alpha_k(1-\alpha_k)}{2}\|\bar{x} - x_{k-1}\|^2 \\
&\quad + \frac{L-\mu}{2}\|z_k - y_k\|^2 + \frac{\epsilon_k}{2}\mathbb{E}_k\left[\|y_k - x_k\|^2\right] - \frac{L-\epsilon_k}{2}\mathbb{E}_k\left[\|z_k - x_k\|^2\right] \\
&\underset{\text{(d)}}{=} \alpha_k F(\bar{x}) + (1-\alpha_k)F(x_{k-1}) - \mathbb{E}_k F(x_k) \\
&\quad + \frac{\alpha_{k-1}^2 L\rho_{k-1}(1-\alpha_k)}{2}\|\bar{x} - v_{k-1}\|^2 + \frac{(\alpha_k - 1)\mu(L\alpha_k - \mu)}{2(L-\mu)}\|x_{k-1} - v_{k-1}\|^2 \\
&\quad + \frac{\epsilon_k}{2}\mathbb{E}_k\left[\|y_k - x_k\|^2\right] - \frac{L-\epsilon_k}{2}\mathbb{E}_k\left[\|z_k - x_k\|^2\right] \\
&\underset{②}{\leq} \alpha_k F(\bar{x}) + (1-\alpha_k)F(x_{k-1}) - \mathbb{E}_k F(x_k) \\
&\quad + \frac{\alpha_{k-1}^2 L\rho_{k-1}(1-\alpha_k)}{2}\|\bar{x} - v_{k-1}\|^2 + \frac{\epsilon_k}{2}\mathbb{E}_k\left[\|y_k - x_k\|^2\right] - \frac{L-\epsilon_k}{2}\mathbb{E}_k\left[\|z_k - x_k\|^2\right] \\
&= (1-\alpha_k)(F(x_{k-1}) - F(\bar{x})) + F(\bar{x}) - \mathbb{E}_k F(x_k) \\
&\quad + \frac{\alpha_{k-1}^2 L\rho_{k-1}(1-\alpha_k)}{2}\|\bar{x} - v_{k-1}\|^2 + \frac{\epsilon_k}{2}\mathbb{E}_k\left[\|y_k - x_k\|^2\right] - \frac{L-\epsilon_k}{2}\mathbb{E}_k\left[\|z_k - x_k\|^2\right] \\
&= (1-\alpha_k)\left(F(x_{k-1}) - F(\bar{x}) + \frac{\alpha_{k-1}^2 L\rho_{k-1}}{2}\|\bar{x} - v_{k-1}\|^2\right) \\
&\quad + F(\bar{x}) - \mathbb{E}_k F(x_k) - \frac{L-\epsilon_k}{2}\mathbb{E}_k\left[\|z_k - x_k\|^2\right] + \frac{\epsilon_k}{2}\mathbb{E}_k\left[\|y_k - x_k\|^2\right] \\
&\underset{③}{=} (1-\alpha_k)\left(F(x_{k-1}) - F(\bar{x}) + \frac{\alpha_{k-1}^2 L\rho_{k-1}}{2}\|\bar{x} - v_{k-1}\|^2\right) \\
&\quad + F(\bar{x}) - \mathbb{E}_k F(x_k) - \frac{\alpha_k^2(L-\epsilon_k)}{2}\mathbb{E}_k\left[\|v_k - \bar{x}\|^2\right] + \frac{\epsilon_k}{2}\mathbb{E}_k\left[\|y_k - x_k\|^2\right].
\end{aligned}
$$

Explanations for ①, ②, ③ are not done yet.

The above produces the following inequality:

$$\begin{aligned}&\mathbb{E}_k F(x_k) - F(\bar{x}) - \frac{\alpha_k^2(L-\epsilon_k)}{2}\mathbb{E}_k\left[\|v_k-\bar{x}\|^2\right]\\&\leq (1-\alpha_k)\left(F(x_{k-1}) - F(\bar{x}) + \frac{\alpha_{k-1}L\rho_{k-1}}{2}\|\bar{x}-v_{k-1}\|^2\right) + \frac{\epsilon_k}{2}\mathbb{E}_k\left[\|y_k-x_k\|^2\right].\end{aligned} \tag{3.1}$$

Proof of (a). Using basic algebra:

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

Proof of (c). Using (b) and some algebra, we can derive:

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

Proof of (d).

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

$$
\begin{aligned}
&\underset{①}{=} -\frac{\mu\alpha_k(1-\alpha_k)}{2}\|\bar{x}-x_{k-1}\|^2 + \frac{L-\mu}{2}\left\|\frac{L\alpha_k-\mu}{L-\mu}(\bar{x}-v_{k-1}) + \frac{\mu(1-\alpha_k)}{L-\mu}(\bar{x}-x_{k-1})\right\|^2 \\
&= -\frac{\mu\alpha_k(1-\alpha_k)}{2}\|\bar{x}-x_{k-1}\|^2 + \frac{(L\alpha_k-\mu)^2}{2(L-\mu)}\|\bar{x}-v_{k-1}\|^2 \\
&\quad + \frac{\mu^2(1-\alpha_k)^2}{2(L-\mu)}\|\bar{x}-x_{k-1}\|^2 + \frac{(L\alpha_k-\mu)\mu(1-\alpha_k)}{L-\mu}\langle\bar{x}-x_{k-1},\bar{x}-v_{k-1}\rangle \\
&= \left(\frac{\mu^2(1-\alpha_k)^2}{2(L-\mu)} - \frac{\mu\alpha_k(1-\alpha_k)}{2}\right)\|\bar{x}-x_{k-1}\|^2 \\
&\quad + \left(\frac{(L\alpha_k-\mu)^2}{2(L-\mu)} - \frac{\alpha_{k-1}^2 L\rho_{k-1}(1-\alpha_k)}{2}\right)\|\bar{x}-v_{k-1}\|^2 \\
&\quad + \frac{\alpha_{k-1}^2 L\rho_{k-1}(1-\alpha_k)}{2}\|\bar{x}-v_{k-1}\|^2 + \frac{(L\alpha_k-\mu)\mu(1-\alpha_k)}{L-\mu}\langle\bar{x}-x_{k-1},\bar{x}-v_{k-1}\rangle \\
&\underset{(a)}{=} \frac{(\alpha_k-1)\mu(L\alpha_k-\mu)}{2(L-\mu)}\|\bar{x}-x_{k-1}\|^2 + \left(\frac{(L\alpha_k-\mu)^2}{2(L-\mu)} - \frac{\alpha_{k-1}^2 L\rho_{k-1}(1-\alpha_k)}{2}\right)\|\bar{x}-v_{k-1}\|^2 \\
&\quad + \frac{\alpha_{k-1}^2 L\rho_{k-1}(1-\alpha_k)}{2}\|\bar{x}-v_{k-1}\|^2 + \frac{(L\alpha_k-\mu)\mu(1-\alpha_k)}{L-\mu}\langle\bar{x}-x_{k-1},\bar{x}-v_{k-1}\rangle \\
&\underset{(c)}{=} \frac{(\alpha_k-1)\mu(L\alpha_k-\mu)}{2(L-\mu)}\|\bar{x}-x_{k-1}\|^2 + \frac{\mu(L\alpha_k-\mu)(\alpha_k-1)}{2(L-\mu)}\|\bar{x}-v_{k-1}\|^2 \\
&\quad + \frac{\alpha_{k-1}^2 L\rho_{k-1}(1-\alpha_k)}{2}\|\bar{x}-v_{k-1}\|^2 + \frac{(L\alpha_k-\mu)\mu(1-\alpha_k)}{L-\mu}\langle\bar{x}-x_{k-1},\bar{x}-v_{k-1}\rangle \\
&= \frac{\alpha_{k-1}^2 L\rho_{k-1}(1-\alpha_k)}{2}\|\bar{x}-v_{k-1}\|^2 \\
&\quad + \frac{(\alpha_k-1)\mu(L\alpha_k-\mu)}{2(L-\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\rho_{k-1}(1-\alpha_k)}{2}\|\bar{x}-v_{k-1}\|^2 + \frac{(\alpha_k-1)\mu(L\alpha_k-\mu)}{2(L-\mu)}\|x_{k-1}-v_{k-1}\|^2.
\end{aligned}
$$

At label ① we used results (i) from Lemma 3.8. □

3.2 Discussion

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.