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

Author *

July 23, 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 fast linear convergence rate, with a square root on the condition number 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 fast 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.

2.1 Basic definitions

{def:pg-opt}

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. Note, it also has $T_\beta(x|f + g) = \operatorname{prox}_{\beta^{-1}g}(x - \beta^{-1}\nabla f(x))$ in optimization literatures.*

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*

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

{def:pg-err-ruler}

Definition 2.6 (Relative proximal gradient error ruler)**.** *Let F satisfies Assumption 2.8. Fix any $x \in \mathbb{R}^n$, suppose that $\tilde{x}$ is an estimated of $T_B(x|F)$. Then the relative proximal gradient error is a set defined as*

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

Remark 2.7. *The definition exists to simplifies the notations for the discussions. When $B > 0$ by strong convexity, the set $\{z : \mathbf{0} \in S_B(z, x|F)\}$ is a singleton, conveniently.*

2.2 Important inequalities

{ass:smooth-plus-nonsmooth}

Assumption 2.8. *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 2.9 (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*

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

{lemma:inex-pg-ineq-proto}

Remark 2.10. *If x, y is out of $\operatorname{dom} F$, the inequality still work by convexity.*

Lemma 2.11 (inexact proximal gradient inequality prototype)**.** *Let $F = f + g$ satisfies Assumption 2.8. Fix any $x \in \mathbb{R}^n$ there exists a $B \geq 0$, $\tilde{x}$ be an estimate of $T_B(x|F)$ such that $D_f(\tilde{x}, x) \leq B/2\|x - \tilde{x}\|^2$. Let $S_B(\tilde{x}, x|F)$ be given by Definition 2.6. Then, for all $z \in \mathbb{R}^n$ and, any $w \in S_B(\tilde{x}, x|F)$ 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. Since $F = f + g$ satisfies Assumption 2.8, for all $B \geq L$, it will be an obvious choice. 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 2.12 (inexact proximal gradient inequality)**.** *Let $F = f + g$ satisfies Definition 2.5 with $L > \mu \geq 0$. Fix arbitrary $x \in \mathbb{R}^n$. Assume the followings:*

(i) *$\tilde{x}$ estimates $T_B(x|F)$.*
(ii) *$B \geq 0$ satisfies $D_f(\tilde{x}, x) \leq B/2\|x-\tilde{x}\|^2$.*
(iii) *Let $S_B(\tilde{x}, x|F)$ be given by Definition 2.6.*

Let $\epsilon \geq 0$ be such that $\|x-\tilde{x}\|\epsilon \geq \operatorname{dist}(\mathbf{0}|S_B(\tilde{x}, x|F))$. 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 2.11 hence, it has for all $z \in \mathbb{R}^n$ the inequality:

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

At ①, we used Cauchy inequality. Since this is true for all $w \in S_B(\tilde{x}, x|F)$, it has:

$$\begin{aligned}\frac{B}{2}\|z-\tilde{x}\|^2 &\leq F(z) - F(\tilde{x}) + \frac{\max(B,L)-\mu}{2}\|z-x\|^2 + \operatorname{dist}(\mathbf{0}|S_B(\tilde{x}, x|F))\|z-\tilde{x}\| \\ &\leq F(z) - F(\tilde{x}) + \frac{\max(B,L)-\mu}{2}\|z-x\|^2 + \epsilon\|x-\tilde{x}\|\|z-\tilde{x}\|.\end{aligned}$$

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 2.13. *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 instead of a random variable to make the algebra possible when deriving the convergence rate of the algorithm.**

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 that gives the analysis of our convergence rate.**

{lemma:pg-stoch-ineq}

Lemma 2.14 (proximal stochastic gradient inequality)**.** *Let $F = f + g$ satisfies Assumption 2.8. Fix any $x, z \in \mathbb{R}^n$. Assume the following*

(i) *$\tilde{\nabla} f(x)$ is a random variable which estimates $\nabla f(x)$, which produces the estimate $\tilde{x}$.*
(ii) *There exists $B \geq 0$ such that $D_f(\tilde{x}, x) \leq B/2\|x - \tilde{x}\|^2$.*

If in addition, there exists some $\epsilon \geq 0$:

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

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 can choose $w = \nabla f(x) - \tilde{\nabla} f(x) \in S_B(\tilde{x}, x|F)$ which is explained in (2.3). Using Lemma 2.11, for any fixed z it has:

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

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

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

□

2.3 discussion

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

{eqn:stoch-grad-epsilon}

$$\frac{\mathbb{E}\left[\left\|\nabla f(x) - \tilde{\nabla} f(x)\right\| \|z-\tilde{x}\|\right]}{\mathbb{E}\left[\|x-\tilde{x}\|\|z-\tilde{x}\|\right]} = \epsilon. \tag{2.4}$$

It depends on both z and x. Let's think about the edge case. When $\mathbb{E}[\|x-\tilde{x}\|\|z-\tilde{x}\|] = 0$, it must be that both $\|x-\tilde{x}\|, \|z-\tilde{x}\|$ are zero, indicating that $\nabla f(x) = \tilde{\nabla} f(x)$ and, $x = \tilde{x} = z$.

Otherwise, we can assume $\mathbb{P}(\|x-\tilde{x}\|\|z-\tilde{x}\|>0)>0$ and, the expectation has:

$$\begin{aligned}&\mathbb{E}\left[\|x-\tilde{x}\|\|z-\tilde{x}\|\right]\\&=\mathbb{E}\left[\|x-\tilde{x}\|\|z-\tilde{x}\| \mid \|x-\tilde{x}\|\|z-\tilde{x}\|>0\right]\mathbb{P}(\|x-\tilde{x}\|\|z-\tilde{x}\|>0).\end{aligned}$$

Let's suppose that $\tilde{\nabla} f(x)$ is a random variable comes from the space: $\Omega(x)$, then it has the following:

$$\begin{aligned}&\mathbb{E}\left[\|x-\tilde{x}\|\|z-\tilde{x}\| \mid \|x-\tilde{x}\|\|z-\tilde{x}\|>0\right]\\&\geq \min_{y\in\Omega(x)}\left\{\|x-\tilde{x}\|: x\neq\tilde{x}=\operatorname*{argmin}_z\left\{g(z)+\langle y,z\rangle+\frac{B}{2}\|z-x\|^2\right\}\right\}\mathbb{P}(\|x-\tilde{x}\|\|z-\tilde{x}\|>0)\mathbb{E}\|z-\tilde{x}\|,\\&\mathbb{E}\left[\left\|\nabla f(x)-\tilde{\nabla} f(x)\right\|\|z-\tilde{x}\|\right]\\&\leq \max_{y\in\Omega(x)}\left\{\|\nabla f(x)-y\|\right\}\mathbb{E}\|z-\tilde{x}\|,\end{aligned}$$

And it would mean:

$$\begin{aligned}\epsilon&=\frac{\mathbb{E}\left[\left\|\nabla f(x)-\tilde{\nabla} f(x)\right\|\|z-\tilde{x}\|\right]}{\mathbb{E}\left[\|x-\tilde{x}\|\|z-\tilde{x}\|\right]}\\&\leq\frac{\max_{y\in\Omega(x)}\|\nabla f(x)-y\|}{\min_{y\in\Omega(x)}\left\{\|x-\tilde{x}\|: x\neq\tilde{x}=\operatorname*{argmin}_z\left\{g(z)+\langle y,z\rangle+\frac{B}{2}\|z-x\|^2\right\}\right\}\mathbb{P}(\|x-\tilde{x}\|\|z-\tilde{x}\|>0)}.\end{aligned}$$

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 has nothing to do with z.

Ok, what about the relation between this error term (2.4) **and the strong growth conditions that is usually found in the literature?** There is no direct equality relationship between the Strong Growth condition of the stochastic gradient which is usually used in the literature, with the condition we have here. Nonetheless, an indirect inequality relation exists. Suppose that the gradient of f satisfies strong growth condition with constant

ρ, so it has $\mathbb{E}\left\|\tilde{\nabla} f(x)\right\|^2 \leq \rho\|\nabla f(x)\|^2$, then $\left\|\nabla f(x) - \tilde{\nabla} f(x)\right\|$ squared has:

$$\begin{aligned}
&\left(\mathbb{E}\left\|\nabla f(x) - \tilde{\nabla} f(x)\right\|\right)^2 \\
&\leq \mathbb{E}\left\|\nabla f(x) - \tilde{\nabla} f(x)\right\|^2 \\
&= \mathbb{E}\left\|\nabla f(x) - \mathbb{E}\tilde{\nabla} f(x)\right\|^2 \\
&= \mathbb{E}\left\|\tilde{\nabla} f(x)\right\|^2 - \|\nabla f(x)\|^2 \leq (\rho - 1)\|\nabla f(x)\|^2.
\end{aligned}$$

The last two equalities, we used the $\mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2$ with X being each of the element of the vector and, the linearity of expectations to apply it individually and them sum it up. The first inequality also uses that equality but $X = \left\|\nabla f(x) - \tilde{\nabla} f(x)\right\|$ instead, and $\mathbb{E}(X - \mathbb{E}X)^2 \geq 0$ always. The last inequality comes from the assumed strong-growth conditions. Strong growth condition does make ϵ smaller if x is close to the set of minimizer and this is the result:

$$\mathbb{E}\left\|\nabla f(x) - \tilde{\nabla} f(x)\right\| \leq \sqrt{\rho - 1}\|\nabla f(x)\|.$$

Keeping the strong growth condition, let's assume that $\tilde{x}$ is generated by stochastic gradient, so that the proximal operator has $g \equiv 0$, then $\tilde{x} = x - L^{-1}\tilde{\nabla} f(x)$ directly. Assuming that relation $m \leq \|z - \tilde{x}\| \leq M$ exists then (2.4) has upper bound

$$\begin{aligned}
&\frac{M}{m}\frac{\mathbb{E}\left\|\nabla f(x) - \tilde{\nabla} f(x)\right\|}{\mathbb{E}\|x - \tilde{x}\|} \\
&\leq \frac{M}{m}\frac{\sqrt{\rho - 1}\|\nabla f(x)\|}{L^{-1}\mathbb{E}\left\|\tilde{\nabla} f(x)\right\|} \\
&\leq \frac{LM}{m}\frac{\sqrt{\rho - 1}\|\nabla f(x)\|}{\min_{y \in \Omega(x)}\{\|y\| : y \neq \mathbf{0}\}\,\mathbb{P}\left(\left\|\tilde{\nabla} f(x)\right\| \neq 0\right)}.
\end{aligned}$$

NOT FINISHED YET. It seems like no pretty relaton exists.

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 (proximal inexact gradient operator with relative error)**.** *Let $F = f + g$ satisfies Assumption 2.8, 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) $\tilde{x} = \mathbf{T}_B(x|F)$ *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 relative error ϵ by:

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

And the inexact output is defined as:

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

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

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.

Definition 3.2 (stochastic proximal gradient operator with relative error)**.** *Let $F = f + g$ satisfies Assumption 2.8. 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) *Any* $B \geq 0$ *such that it satisfies* $D_f(\tilde{x}, x) \leq B/2\|x - \tilde{x}\|^2$.
(iii) ϵ *is an relative error, determined by* $z, x, \tilde{\nabla} f(x)$ *and, defined immediately below.*

The relative error ϵ is defined as:

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

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

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

Remark 3.3. *For this definition of the stochastic proximal gradient operator, Lemma 2.14 is applicable.*

{def:SNAG}

Definition 3.4 (inexact/stochastic SNAG)**.** *Suppose that $F = f + g$ satisfies Assumption 2.8. 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:*

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

The following definition gives a momentum sequence where it makes the derivation of the convergence rate easier. The sequence is powerful, instead of an equality relation, the sequence can instead satisfies $\alpha_k \in (\mu/L, 1)$.

{def:relax-momen-seq}

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 $\alpha_k \in (\mu/L, 1)$.*

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.*

The following lemma defines the conditions required for the momentum sequence to be compatible with convergence claims.

{lemma:seq-properties}

Lemma 3.7 (relaxed momentum sequence conditions)**.** *Let $(\alpha_k)_{k\geq 0}$, L, μ to be two constant such that it has $L > \mu \geq 0$. If we assume that the sequence has for all $k \geq 0 : \alpha_k \geq \mu/L$, then we can define a positive sequence*

$$(\forall k \geq 1) : \rho_{k-1} = \frac{\alpha_k(\alpha_k - \mu/L)}{(1-\alpha_k)\alpha_{k-1}^2}.$$

And under this relationship, the followings are true inductively:

(i) *In general, for all $\rho_{k-1} \geq 0$, it has $0 \leq \alpha_k \leq \min\left(1, |\rho\alpha^2 - q| + \sqrt{\rho}\alpha\right)$.*
(ii) *If $\alpha_{k-1} \geq \mu/L$, then $\alpha_k \geq \mu/L$ too.*
(iii) *If $\alpha_{k-1} > \mu/L > 0$, then $\alpha_k \in (\mu/L, 1)$ for all $\rho_{k-1} > 0$.*

Proof. **Proof of item (i), (iii)**. With $\rho_{k-1}(1-\alpha_k)\alpha_{k-1}^2 = \alpha_k(\alpha_k - \mu/L)$ it gives:

$$\alpha_k = \frac{1}{2}\left(\frac{\mu}{L} - \rho_{k-1}\alpha_{k-1}^2 + \sqrt{\left(\rho_{k-1}\alpha_{k-1}^2 - \frac{\mu}{L}\right)^2 + 4\rho_{k-1}\alpha_{k-1}^2}\right) > 0.$$

Focusing exclusively on the RHS, we omit the subscript and write α_{k-1}, ρ_{k-1} as α, ρ. We also just write $q = \mu/L$. By definition, we have $L > \mu \geq 0$ hence, $q \in [0, 1)$. We also note that the quantity $(r\alpha^2 + q)^2 + 4\rho\alpha$ is clearly > 0. We will show that there is an upper bound for the RHS in terms of ρ, q. Completing the square should give

$$\begin{aligned}
0 &\leq \left(\rho\alpha^2 - q\right)^2 + 4\rho\alpha^2 \\
&= \rho^2\alpha^4 + q^2 - 2\rho\alpha^2 q + 4\rho\alpha^2 \\
&= \rho^2\alpha^4 + 2\rho(2-q)\alpha^2 + \rho^2(2-q)^2 - \rho^2(2-q)^2 + q^2 \\
&= (\rho\alpha^2 + \rho(2-q))^2 - \rho^2(2-q)^2 + q^2 \\
&= \rho^2(\alpha^2 + 2 - q)^2 + q^2 - \rho^2(2-q)^2 \\
&\leq \rho^2(\alpha^2 + 2 - q)^2 + \max(0, q^2 - \rho^2(2-q)^2).
\end{aligned}$$

The above is non-negative, taking the square root it has:

$$\begin{aligned}
\sqrt{\left(\rho\alpha^2 - q\right)^2 + 4\rho\alpha^2} &\leq \rho|\alpha^2 + 2 - q| + \sqrt{\max(0, q^2 - \rho^2(2-q)^2)} \\
&\leq \rho|\alpha^2 + 2 - q|
\end{aligned}$$

But look, if we directly apply the $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ it has another upper bound which is:

{ineq:seq-properties-pr3}

$$\sqrt{\left(\rho\alpha^2 - q\right)^2 + 4\rho\alpha^2} \leq |\rho\alpha^2 - q| + 2\sqrt{\rho}\alpha. \tag{3.1}$$

Both upper bounds apply and hence we have:

$$\begin{aligned}
\alpha_k &\leq \frac{1}{2}\left(q - \rho\alpha^2 + \min\left(\rho|\alpha^2 + 2 - q|, |\rho\alpha^2 - q| + 2\sqrt{\rho}\alpha\right)\right) \\
&= \frac{1}{2}\left(q - \rho\alpha^2 + \min\left(\rho\alpha^2 + (2 - q), |\rho\alpha^2 - q| + 2\sqrt{\rho}\alpha\right)\right) \\
&= \frac{1}{2}\left(\min\left(2, |\rho\alpha^2 - q| + 2\sqrt{\rho}\alpha + q - \rho\alpha^2\right)\right) \\
&= \min\left(1, (1/2)|\rho\alpha^2 - q| + \sqrt{\rho}\alpha + (1/2)(q - \rho\alpha^2)\right) \\
&\leq \min\left(1, |\rho\alpha^2 - q| + \sqrt{\rho}\alpha\right).
\end{aligned}$$

Proof of (ii). To see the lower bound, we assume that $\alpha_k \geq \mu/L$, denote μ/L by q, we have the following:

{ineq:seq-properties-pr5}

$$\alpha_k - q \underset{①}{\geq} \alpha_k(\alpha_k - q) = \rho_{k-1}(1 - \alpha_k)\alpha_{k-1}^2 \underset{②}{\geq} \rho_{k-1}(1 - \alpha_k)q^2 \geq 0. \tag{3.2}$$

At ①, we used the fact that $\alpha_k \leq 1$ and $\alpha_k \geq 0$ from previous results. At ②, we used the assumption that $\alpha_{k-1} \geq q \geq 0$. The last inequality is true because is $\rho_{k-1} \geq 0, \alpha_k \leq 1$.

Proof of (iii). In addition, if $\alpha > \mu/L > 0$, then the inequality in (3.1) is strict, then it would instead make $\alpha_k < 1$. It will also make the inequality chain in (3.2) strict. Therefore, Item (iii) is true. □

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:

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

□

Remark the following definitions

Definition 3.9 (conditional expectations)**.** *Given probability space* $(\Omega, \mathcal{F}_0, \mathbb{P})$ *and,* $\mathcal{F} \subseteq \mathcal{F}_0$*, and a random variable* $X \in \mathcal{F}$ *such that* $\mathbb{E}|X| < \infty$*. We define the conditional expectation of* X *given* $\mathcal{F}$*,* $\mathbb{E}[X|\mathcal{F}]$ *to be any random variable* Y *that has*

(i) $Y \in \mathcal{F}$*, i.e:* Y *is* $\mathcal{F}$ *measurable.*
(ii) *For all* $A \in \mathcal{F}$*,* $\int_A X d\mathbb{P} = \int_A Y d\mathbb{P}$*.*

Remark 3.10.

CITATIONS HERE NEEDED.

{thm:snag-descent}

The conditional expectation exists and it's unique. The following theorem given an inequality characterizing a descent relation for the SNAG algorithm.

Theorem 3.11 (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* $(\alpha_k)_{k\geq 0}$ *is given as in Lemma 3.7.*

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:*

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

In the edge case of $k = 0$*, it has:*

$$\mathbb{E}_0 F(x_0) - F(\bar{x}) + \frac{L-\epsilon}{2}\mathbb{E}_0\left[\|\bar{x} - x_0\|^2\right] \leq \frac{L-\mu}{2}\|\bar{x} - v_{-1}\|^2 + \frac{\epsilon_0}{2}\mathbb{E}_0\left[\|v_{-1} - x_0\|^2\right].$$

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 $k \geq 1$: $\rho_{k-1}(1-\alpha_k)\alpha_{k-1}^2 = \alpha_k(\alpha_k - \mu/L)$, hence Lemma 3.7 applies.

(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 2.14 with $x = y_k, z = z_k, \tilde{x} = x_k$ and, $B = L$ then it means:

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

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

At ①, we used Lemma 2.9. For ②, we used Lemma 3.7 (i) because of what is assumed in (b) so, the sequence has $\alpha_k \leq 1$, therefore $1-\alpha_k \leq 0$. Next, $\alpha_0 = 1 > \mu/L$, Lemma 3.7 (ii) applies too, hence $L\alpha_k - \mu \geq 0$. Since $L > \mu$ always, the coefficient on $\|x_{k-1} - v_{k-1}\|^2$ is always ≤ 0, therefore the term can be removed to keep the inequality. At ③, we used Lemma 3.8 (ii). Therefore, at the end the chain of inequalities from above produced the following inequality:

$$
\begin{aligned}
&\mathbb{E}_k F(x_k) - F(\bar{x}) + \frac{\alpha_k^2(L-\epsilon_k)}{2}\mathbb{E}_k\left[\|v_k - \bar{x}\|^2\right] \\
&\leq (1-\alpha_k)\left(F(x_{k-1}) - F(\bar{x}) + \frac{\alpha_{k-1}^2 L\rho_{k-1}}{2}\|\bar{x} - v_{k-1}\|^2\right) + \frac{\epsilon_k}{2}\mathbb{E}_k\left[\|y_k - x_k\|^2\right].
\end{aligned} \tag{3.3}
$$

Next, we handle the Edge case of $k = 0$. The base case has $\alpha_0 = 1$, and $v_{-1} = x_{-1}$. Then, the following consequences will be immediate. Fom Definition 3.4, $\alpha_0 = 1$ makes $\tau_0 = 0$, so $y_0 = v_{-1}$. Followed by it, $x_0 = \tilde{\mathbf{T}}_L(y_0|F) = \widetilde{\mathbf{T}}_L(v_{-1}|F)$. Therefore, we can apply Lemma 2.14 with $z = \bar{x}$, $\tilde{x} = x_0$, $x = v_{-1}$, and $\epsilon = \epsilon_0$ it yields:

$$
\begin{aligned}
0 &\leq F(z) - \mathbb{E}_0 F(\tilde{x}) + \frac{L-\mu}{2}\|z - x\|^2 - \frac{L-\epsilon}{2}\mathbb{E}_0\left[\|z - \tilde{x}\|^2\right] + \frac{\epsilon}{2}\mathbb{E}_0\left[\|x - \tilde{x}\|^2\right] \\
&= F(\bar{x}) - \mathbb{E}_0 F(x_0) + \frac{L-\mu}{2}\|\bar{x} - v_{-1}\|^2 - \frac{L-\epsilon}{2}\mathbb{E}_0\left[\|\bar{x} - x_0\|^2\right] + \frac{\epsilon}{2}\mathbb{E}_0\left[\|v_{-1} - x_0\|^2\right].
\end{aligned}
$$

Proof of (a). Using basic algebra:

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

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).

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

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

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

Obtaining the convergence results from the lemma requires a recursive relation in expectation. The following theorem gives necessary results to derive the convergence rate of the algorithm in expectations.

{thm:snag-generic-cnvg}

Theorem 3.12 (SNAG generic convergence rate)**.** *Suppose that the sequence $(y_k, x_k, v_k)_{k\geq 0}$ is generated by algorithms satisfying Definition 3.4. Assume that the momentum sequence $(\alpha_k)_{k\geq 0}$, stochastic gradient $\tilde{\nabla} f$, and v_{-1}, x_{-1} as in Theorem 3.11. For all $\bar{x} \in \mathbb{R}^n$, for all $k \geq 1$, the convergence in expectation, can be compactly written as:*

$$
\mathbb{E}\Phi_k \leq \left(\prod_{i=1}^{k}\lambda_i\right)\mathbb{E}\Phi_0 + \sum_{j=0}^{k-1}\left(\prod_{i=k-j}^{k}\lambda_i\right)\mathbb{E}R_{k-j-1} + \mathbb{E}R_k.
$$

Where:

(i) $\Phi_k := F(x_k) - F(\bar{x}) + \alpha_k^2(L-\epsilon_k)/2\|v_k-\bar{x}\|^2$,
(ii) $R_k := \frac{\epsilon_k}{2}\|y_k - x_k\|^2$,
(iii) $\lambda_k := (1-\alpha_k)\max(1, L\rho_{k-1}/(L-\epsilon_{k-1}))$.

Proof. Denote $\mathbb{E}_k$ to be the conditional expectation based on random variables $\tilde{\nabla} f(y_i)$ for $i = 1, \ldots, k-1$. The base case has $\mathbb{E}_0$ which is the overall expectation. Fix any $\bar{x} \in \mathbb{R}^n$ throughout. Consider any $k \geq 1$. Under the new notations, the relations from Theorem 3.11

yields

$$
\begin{aligned}
&\mathbb{E}_k\Phi_k \\
&= \mathbb{E}_k F(x_k) - F(\bar{x}) + \frac{\alpha_k(L-\epsilon_k)}{2}\mathbb{E}_k\|v_k - \bar{x}\|^2 \\
&\underset{①}{\leq} (1-\alpha_k)\left(F(x_{k-1}) - F(\bar{x}) + \frac{\alpha_{k-1}^2 L\rho_{k-1}}{2}\|\bar{x} - v_{k-1}\|^2\right) + \frac{\epsilon_k}{2}\mathbb{E}_k\|y_k - x_k\|^2 \\
&= (1-\alpha_k)\left(F(x_{k-1}) - F(\bar{x}) + \frac{\alpha_{k-1}^2 L\rho_{k-1}}{2}\left(\frac{L-\epsilon_{k-1}}{L}\right)\left(\frac{L}{L-\epsilon_{k-1}}\right)\|\bar{x} - v_{k-1}\|^2\right) + \mathbb{E}_k R_k \\
&= (1-\alpha_k)\left(F(x_{k-1}) - F(\bar{x}) + \frac{\alpha_{k-1}^2 (L-\epsilon_{k-1})}{2}\left(\frac{L\rho_{k-1}}{L-\epsilon_{k-1}}\right)\|\bar{x} - v_{k-1}\|^2\right) + \mathbb{E}_k R_k \\
&= (1-\alpha_k)\max\left(1, \frac{L\rho_{k-1}}{L-\epsilon_{k-1}}\right)\left(F(x_{k-1}) - F(\bar{x}) + \frac{\alpha_{k-1}^2 (L-\epsilon_{k-1})}{2}\|\bar{x} - v_{k-1}\|^2\right) + \mathbb{E}_k R_k \\
&= \lambda_k\Phi_{k-1} + \mathbb{E}_k R_k.
\end{aligned}
$$

At ①, we used Theorem 3.11. Taking the full expectation, which is $\mathbb{E}_0$ on both sides of the inequality above it yields

{ineq:snag-descent-expe-pitem-1}

$$\mathbb{E}_0\Phi_k \leq \lambda_{k+1}\mathbb{E}_0\Phi_k + \mathbb{E}_0 R_{k+1}. \tag{3.4}$$

Let $k \geq 1$, We are now in position to verify the following inductive hypothesis:

$$\mathbb{E}_0\Phi_k \leq \left(\prod_{i=1}^{k}\lambda_i\right)\mathbb{E}_0\Phi_0 + \sum_{j=0}^{k-1}\left(\prod_{i=k-j}^{k}\lambda_i\right)\mathbb{E}_0 R_{k-j-1} + \mathbb{E}_0 R_k. \tag{IH}$$

Use Inequality 3.4 with k as $k+1$, the inductive proof follows:

$$
\begin{aligned}
\mathbb{E}_0\Phi_{k+1} &\leq \lambda_{k+1}\mathbb{E}_0\Phi_k + \mathbb{E}_0 R_{k+1} \\
&\underset{\text{(IH)}}{\leq} \lambda_{k+1}\left(\left(\prod_{i=1}^{k}\lambda_i\right)\mathbb{E}_0\Phi_0 + \sum_{j=0}^{k-1}\left(\prod_{i=k-j}^{k}\lambda_i\right)\mathbb{E}_0 R_{k-j-1} + \mathbb{E}_0 R_k\right) + \mathbb{E}_0 R_{k+1} \\
&= \left(\prod_{i=1}^{k+1}\lambda_i\right)\mathbb{E}_0\Phi_0 + \sum_{j=0}^{k-1}\left(\prod_{i=k-j}^{k+1}\lambda_i\right)\mathbb{E}_0 R_{k-j-1} + \lambda_{k+1}\mathbb{E}_0 R_k + \mathbb{E}_0 R_{k+1} \\
&= \left(\prod_{i=1}^{k+1}\lambda_i\right)\mathbb{E}_0\Phi_0 + \sum_{j=1}^{k}\left(\prod_{i=k-j}^{k+1}\lambda_i\right)\mathbb{E}_0 R_{k-j} + \sum_{j=0}^{0}\prod_{i=k+1-j}^{k+1}\lambda_{k+1}\mathbb{E}_0 R_k + \mathbb{E}_0 R_{k+1} \\
&= \left(\prod_{i=1}^{k+1}\lambda_i\right)\mathbb{E}_0\Phi_0 + \sum_{j=0}^{k}\left(\prod_{i=k-j}^{k+1}\lambda_i\right)\mathbb{E}_0 R_{k-j} + \mathbb{E}_0 R_{k+1}.
\end{aligned}
$$

Therefore, the inductive hypothesis holds. □

3.2 Discussion of error schedules

Let's talk about Theorem 3.12 here. With the compact representations, we can make see what is happening here. The inequality of the convergence consists of several quantities:

(i) Φ_0, the base case and a bounded quantity, depends on initial iterate v_{-1}, x_{-1}.
(ii) λ_k, a sequence controlling the convergence, involving relaxation parameters ρ_{k-1} and, the error ϵ_k.
(iii) R_k, an error term related to ϵ_k and, $\|y_k - x_k\|^2$, it's none trivial to group with the other terms.

Without necessarily introducing the interpolation hypothesis, we have linear "convergence" with an extra additive term under mild assumption. The following propositions simply stated the sufficient conditions on Φ_k, λ_k, R_k for the inequality to yield a linear convergence rate with additive error, nothing more.

{prop:snag-kinda-converge}

Proposition 3.13 (SNAG kinda converges linearly)**.** *Suppose that the sequence $(y_k, x_k, v_k)_{k\geq 0}$ is generated by algorithms satisfiying 3.4. Suppose that that Φ_k, λ_k, R_k are as given in Theorem 3.12. Define*

(i) $\Lambda = \sup_{i\in\mathbb{N}} \lambda_i$.
(ii) *An upper bound $\bar{R} \geq \mathbb{E}R_k$ exists for all $k \geq 0$.*
(iii) *An upper bound $\overline{\epsilon} = \sup_{i\in\mathbb{N}\cup\{0\}} \epsilon_i$ exists.*

If in addition, we have $\overline{\epsilon} < \sqrt{\mu L}$, then it satisfies for all $k \geq 1$ the following inequality:

$$\mathbb{E}\Phi_k \leq \Lambda^k \mathbb{E}\Phi_0 + \frac{\overline{R}}{1-\Lambda}.$$

Where

$$\Lambda = \left(1 - \sqrt{\frac{\mu}{L}}\right)\left(\frac{L}{L-\overline{\epsilon}}\right).$$

Proof. We can prove it with the following intermediate results:

(a) For all $k \geq 1$, $\alpha_k = \sqrt{\mu/L}$ satisfies recurrences $(1-\alpha_k)\alpha_{k-1}^2 = \alpha_k(\alpha_k - \mu/L)$ where $\rho_k = 1$ for all $k \geq 1$, and $\rho_0 = \mu/L$.
(b) When $(\epsilon_k)_{k\geq 0}$ satisfies $\epsilon_k \leq \bar{\epsilon} = \sup_{i\in\mathbb{N}\cup\{0\}} \epsilon_i < \sqrt{\mu L}$, it has $\lambda_k \in (0,1)$ for all $k \geq 1$, with $\Lambda = \sup_{i\in\mathbb{N}} \lambda_i < 1$.

These results will be proved at the end. Consider starting with the inequality proved in

Theorem 3.11:

$$\begin{aligned}
\mathbb{E}\Phi_k &\leq \left(\prod_{i=1}^{k}\lambda_i\right)\mathbb{E}\Phi_0 + \sum_{j=1}^{k-1}\left(\prod_{j=k-j}^{k-1}\lambda_i\right)\mathbb{E}R_{k-j-1} + \mathbb{E}R_k \\
&\underset{①}{\leq} \Lambda^k\mathbb{E}\Phi_0 + \sum_{j=0}^{k-1}\Lambda^{j+1}\mathbb{E}R_{k-j-1} + \mathbb{E}R_k \\
&= \Lambda^k\mathbb{E}\Phi_0 + \sum_{j=1}^{k}\Lambda^{j}\mathbb{E}R_{k-j} + \mathbb{E}R_k \\
&= \Lambda^k\mathbb{E}\Phi_0 + \sum_{j=0}^{k}\Lambda^{j}\mathbb{E}R_{k-j} \\
&\underset{②}{\leq} \Lambda^k\mathbb{E}\Phi_0 + \overline{R}\sum_{j=0}^{k}\Lambda^{j} \\
&\underset{③}{\leq} \Lambda^k\mathbb{E}\Phi_0 + \frac{\overline{R}}{1-\Lambda}.
\end{aligned}$$

At ① we used that $\Lambda \geq \lambda_k$. At ② we used that $\overline{R} \geq \mathbb{E}R_k$ for all $k \geq 0$. At ③ we used the results from (b) that $\lambda_k \leq \Lambda$ and $\Lambda < 1$, hence it has the convergence of the geometric series.

Proof of (a) When $L > \mu > 0$, so the objective F is strongly convex, the usual $(1-\alpha_k)\alpha_{k-1}^2 = \alpha_k(\alpha_k - \mu/L)$ recurrence relations admits a constant solution $\alpha_k = \sqrt{\mu/L}$ for all $k \geq 1$. As presented in Theorem 3.11, the momentum sequence $(\alpha_k)_{k\geq 0}$ has a relaxation parameter $(\rho_k)_{k\geq 1}$. In this case, it means that $\rho_{k-1} = 1$ for all $k \geq 1$. Since when $k = 0$, $\alpha_0 = 1$, it makes an exception on ρ_0, which is:

$$\rho_0 = \frac{\sqrt{\mu/L}\left(\sqrt{\mu/L} - \mu/L\right)}{1 - \sqrt{\mu/L}} = \frac{\mu}{L}.$$

Proof of (b). If $\sup_{k\in\mathbb{N}\cup\{0\}} \epsilon_k < \sqrt{\mu L}$ for all $k \geq 0$, by the definition of λ_k, it has:

$$\begin{aligned}
&\sup_{k\in\mathbb{N}} \lambda_k \\
&= \sup_{k\in\mathbb{N}} \left(1 - \sqrt{\frac{\mu}{L}}\right)\left(\frac{L\rho_{k-1}}{L - \epsilon_{k-1}}\right) \\
&\underset{①}{\leq} \left(1 - \sqrt{\frac{\mu}{L}}\right)\left(\frac{L}{L - \sup_{k\in\mathbb{N}} \epsilon_{k-1}}\right) \\
&< \left(1 - \sqrt{\frac{\mu}{L}}\right)\left(\frac{L}{1 - \sqrt{\mu L}}\right) = 1.
\end{aligned}$$

At ①, we used the fact that $\rho_k = 1$ for all $k \geq 1$ and,$\rho_k = \mu/L$ for $k = 0$, therefore, it $(\forall k \geq 0)\ \rho_k \leq 1$. So $\Lambda < 1$.

□

4 Interpolation helps but, it's very unclear if it's sufficient

Under the frameworks of our analysis, the interpolation assumption helps with controlling the errors, but it's unclear to what extent, it will help us to obtain a convergence rate faster than what is already in the literatures.

4.1 The no error case

4.2 The degenerate case

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.