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

Author *

July 14, 2025

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

Abstract

This file is for communication purposes between collaborators.

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

1 Introduction

Does stochastic accelerated Nesterov's acceleration (SNAG) produces accelerated convergence rate (or, any type of convergence) when the Interpolation Hypothesis is true? [1] Previously we got some results, but unfortunately it was incorrect the mistake is diffcult to recover.

However, 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.

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

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.

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.*

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*

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

{lemma:inex-pg-ineq}

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

Lemma 2.9 (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 $\tilde{x}$, characterized by:*

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

Then, it would satisfy for all $z \in \mathbb{R}^n$ 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. 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 $z \in \mathbb{R}^n$:

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

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

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

At (b) we used the Cauchy Inequality and, the assumption $\|w\| \leq \epsilon\|x-\tilde{x}\|$. 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 2.10. *Usually in practice, the precise value of $F(\tilde{x})$ is never known, therefore, B cannot be easily determined via $D_f(\tilde{x}, x)$. Hence, in this case we can only choose $B \geq L$ which gives:*

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

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:

$$\begin{aligned} \mathbf{0} &\in \partial g(\tilde{x}) + \tilde{\nabla} f(x) + B(\tilde{x} - x) \\ &= \partial g(\tilde{x}) + \tilde{\nabla} f(x) - \nabla f(x) + \nabla f(x) + B(\tilde{x} - x) \\ \iff \nabla f(x) - \tilde{\nabla} f(x) &\in \partial g(\tilde{x}) + \nabla f(x) + B(\tilde{x} - x). \end{aligned}$$

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 $\nabla f(x)$.

3 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 $\tilde{\nabla} f(x)$ is an estimate of the true gradient $\nabla f(x)$ with relative error $\epsilon \geq 0$. Then, the inexact proximal gradient operator is defined by:*

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

And it satisfies

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

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

{def:SNAG}

Definition 3.2 (stochastic proximal gradient operator with relative error)**.**

Definition 3.3 (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:*

$$\begin{aligned}
\tau_k &= L_k(1-\alpha_k)\left(L_k\alpha_k - \sigma^{(k)}\right)^{-1}, \\
y_k &= (1+\tau_k)^{-1}v_{k-1} + \tau_k(1+\tau_k)^{-1}x_{k-1}, \\
x_k &= T_L^{(\epsilon_k)}(y_k|F_{I_k}), \\
v_k &= x_{k-1} + \alpha_k^{-1}(x_k - x_{k-1}).
\end{aligned}$$

{lemma:snag-identities}

The following lemma states two important relationships on the iterates generated by Definition 3.3.

Lemma 3.4 (SNAG identities)**.**

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.