Linear Convergence of Accelerated Gradient without Restart

Author 1 Name, Author 2 Name *

March 2, 2020

This paper is currently in draft mode. Check source to change options. [1]

Abstract

This is still a note for a draft so no abstract.

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

1 Introduction

Notations. Unless specified, our ambient space is $\mathbb{R}^n$ with Euclidean norm $\|\cdot\|$. Let $C \subseteq \mathbb{R}^n$, $\Pi_C(\cdot)$ denotes the projection onto the set C, i.e: the closest point in C to another point in $\mathbb{R}^n$. For a function of $F = f + g$, and a $B \geq 0$ where f is $\mathcal{C}^1$ differentiable, and g is l.s.c, we consider the proximal gradient operator:

$$\begin{aligned} T_B(x) &= \underset{z}{\operatorname{argmin}} \left\{ g(z) + f(x) + \langle \nabla f(x), z - x \rangle + \frac{B}{2}\|x - z\|^2 \right\} \\ &= \operatorname{prox}_{B^{-1}g}(x - B^{-1}\nabla f(x)). \end{aligned}$$

We also define the gradient mapping operator $\mathcal{G}_B(x) = B^{-1}(x - T_B(x))$.

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

2 Precursors materials for our proofs of linear convergence

The following two definitions defines the accelerated proximal gradient algorithm.

{def:st-apg} **Definition 2.1 (similar triangle form of accelerated proximal gradient)** *The definition is about* $((\alpha_k)_{k\geq 0}, (q_k)_{k\geq 0}, (B_k)_{k\geq 0}, (y_k)_{k\geq 0}, (x_k)_{k\geq -1}, (v_k)_{k\geq -1})$. *These sequences satisfy:*

(i) $x_{-1}, y_{-1} \in \mathbb{R}^n$ *are arbitrary initial condition of the algorithm;*
(ii) $(q_k)_{k\geq 1}$ *be a sequence such that* $q_k \in [0, 1)$ *for all* $k \geq 1$*;*
(iii) $(\alpha_k)_{k\geq 1}$ *be a sequence such that* $\alpha_0 \in (0, 1]$*, and for all* $k \geq 1$ *it has* $\alpha_k \in (q_k, 1)$*;*
(iv) $(B_k)_{k\geq 0}$ *has* $B_k \geq 0$.

Then an algorithm satisfies the similar triangle form of Nesterov's accelerated gradient if it generates iterates $(y_k, x_k, v_k)_{k\geq 1}$ *such that for all* $k \geq 0$*:*

$$
\begin{aligned}
y_k &= \left(\frac{\alpha_k - q_k}{1 - q_k}\right) v_{k-1} + \left(\frac{1 - \alpha_k}{1 - q_k}\right) x_{k-1}, \\
x_k &= T_{L_k}(y_k), D_f(x_k, y_k) \leq \frac{B_k}{2}\|x_k - y_k\|^2, \\
v_k &= x_{k-1} + \alpha_k^{-1}(x_k - x_{k-1}).
\end{aligned}
$$

{def:rlx-momentum-seq} **Definition 2.2 (relaxed momentum sequence)** *The following definition is about sequences* $((\alpha_k)_{k\geq 0}, (q_k)_{k\geq 0}, (\rho_k)_{k\geq 0})$. *Let*

(i) $(q_k)_{k\geq 0}$ *is a sequence such that* $q_k \in [0, 1)$ *for all* $k \geq 0$*;*
(ii) $(\alpha_k)_{k\geq 0}$ *be such that* $\alpha_0 \in (0, 1]$*, and for all* $k \geq 1$ *it has* $\alpha_k \in (q_k, 1)$*;*
(iii) $(\rho_k)_{k\geq 0}$ *is a strictly positive sequence for all* $k \geq 1$.

The sequences q_k, α_k *are considered relaxed momentum sequence if for all* $k \geq 1$ *it satisfies the relation that:*

$$
\rho_{k-1} = \frac{\alpha_k(\alpha_k - q_k)}{(1 - \alpha_k)\alpha_{k-1}^2}.
$$

{def:pg-gap} **Definition 2.3 (proximal gradient gap)** *Let* $F = f + g$ *where* f *is* L *Lipschitz smooth and* g *is convex. Then the proximal gradient mapping* $T_B(x) = \text{prox}_{B^{-1}g}(x - B^{-1}\nabla f(x))$ *is a singleton, which as domain on* $\mathbb{R}^n$. *Let* μ, B *be parameters such that* $B > \mu \geq 0$. *We define the proximal gradient gap* $\mathcal{E}(z, y, \mu)$ *is a* $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ *mapping:*

$$
\mathcal{E}(z, y, \mu, B) := F(z) - F(T_B(y)) - \langle B(y - T_B(y)), z - y\rangle - \frac{\mu}{2}\|z - y\|^2 - \frac{B}{2}\|y - T_B(y)\|^2.
$$

Remark 2.4 This expression is the same as the proximal gradient inequality.

3 Deriving the convergence rate

To derive the convergence rate of algorithm satisfying Definition 2.1, 2.2, we leverage Definition 2.3.

{ass:for-cnvg}

Assumption 3.1 (for convergence) The following assumption is about $(F, f, g, \mathcal{E}, \mu, L)$, it is the configuration needed to derive the convergence rate of algorithms that satisfy Definition 2.1. There exists $B > \mu \geq 0$ such that the following are true.

(i) Let $F = f + g$ where f is L Lipschitz smooth and, g is closed convex and proper.
(ii) Assume that $X^+ = \underset{x}{\operatorname{argmin}}\{f(x) + g(x)\}$ has $X^+ \neq \emptyset$.
(iii) $\forall z \in \mathbb{R}^n$ it has $\mathcal{E}(\Pi_{X^+}(y), y, \mu, B) \geq 0$.
(iv) For all $z, y \in \mathbb{R}^n$, it has $\mathcal{E}(z, y, \mu, B) + \frac{\mu}{2}\|z - y\|^2 \geq 0$.

{lemma:st-iterates-alt-form-part1}

Note that, if the function is convex, all conditions are satisfies for $\mu = 0$.

Lemma 3.2 (equivalent representations of the iterates part I) *Suppose that the sequences of $\alpha_k, q_k, y_k, v_k, x_k, B_k$ satisfy the similar triangle form, then for all $k \geq 0$ the iterates v_k admits the following equivalent representations:*

$$
\begin{aligned}
v_k &= x_{k-1} + \alpha_k^{-1}(x_k - x_{k-1}) \\
&= v_{k-1} + \alpha_k^{-1} q_k (y_k - v_{k-1}) - \alpha_k^{-1} B_k^{-1} \mathcal{G}_{B_k}(y_k).
\end{aligned}
$$

Proof. Consider all $k \geq 1$. The relations is direct, immediately from the update rule in Definition 2.1 of y_k we have

(a) $(\alpha_k - 1)x_{k-1} = (\alpha_k - q_k)v_{k-1} - (1 - q_k)y_k$.
(b) $x_k = y_k - B_k^{-1}\mathcal{G}_{B_k}(y_k)$.

$$
\begin{aligned}
v_k &= x_{k-1} + \alpha_k^{-1}(x_k - x_{k-1}) \\
&= (1 - \alpha_k^{-1})x_{k-1} + \alpha_k^{-1}x_k \\
&= \alpha_k^{-1}(\alpha_k - 1)x_{k-1} + \alpha_k^{-1}x_k \\
&\underset{\text{(a)}}{=} \alpha_k^{-1}(\alpha_k - q_k)v_{k-1} - \alpha_k^{-1}(1 - q_k)y_k + \alpha_k^{-1}x_k \\
&\underset{\text{(b)}}{=} (1 - \alpha_k^{-1}q_k)v_{k-1} - (\alpha_k^{-1} - \alpha_k^{-1}q_k)y_k + \alpha_k^{-1}(y_k - B_k^{-1}\mathcal{G}_{B_k}(y_k)). \\
&= (1 - \alpha_k^{-1}q_k)v_{k-1} + \alpha_k^{-1}q_k y_k - \alpha_k^{-1}B_k^{-1}\mathcal{G}_{B_k}(y_k) \\
&= v_{k-1} + \alpha_k^{-1}q_k(y_k - v_{k-1}) - \alpha_k^{-1}B_k^{-1}\mathcal{G}_{B_k}(y_k).
\end{aligned}
$$

■

{lemma:st-iterates-alt-form-part2} **Lemma 3.3 (equivalent representations of the iterates part II)** *Suppose the sequences of $\alpha_k, q_k, y_k, v_k, x_k, B_k$ satisfy the similar triangle form, then for all $k \geq 0$ the iterates v_k admits the following equivalent representations:*

$$\begin{aligned} y_k &= x_{k-1} + (1-q_k)^{-1}(\alpha_{k-1}^{-1} - 1)(\alpha_k - q_k)(x_{k-1} - x_{k-2}) \\ &= \left(\frac{\alpha_k - q_k}{1-q_k}\right) v_{k-1} + \left(\frac{1-\alpha_k}{1-q_k}\right) x_{k-1}. \end{aligned}$$

Proof. For all $k \geq 1$, from the update rules in Definition 2.1:

$$\begin{aligned} (1-q_k)^{-1} y_k &= (\alpha_k - q_k) v_{k-1} + (1-\alpha_k) x_{k-1} \\ &= (\alpha_k - q_k)\left(x_{k-2} + \alpha_{k-1}^{-1}(x_{k-1} - x_{k-2})\right) + (1-\alpha_k)x_{k-1} \\ &= (\alpha_k - q_k)x_{k-2} + \alpha_{k-1}^{-1}(x_{k-1} - x_{k-2}) + (1-\alpha_k)x_{k-1} \\ &= (\alpha_k - q_k)(1 - \alpha_{k-1}^{-1})x_{k-2} + \left(\frac{\alpha_k - q_k}{\alpha_{k-1}} + 1 - \alpha_k\right) x_{k-1}. \end{aligned}$$

Multiply $(1-q_k)$ on both sides yield:

$$\begin{aligned} y_k &= \frac{(\alpha_k - q_k)(1-\alpha_{k-1}^{-1})}{1-q_k} x_{k-2} + \left(\frac{\alpha_k - q_k}{\alpha_{k-1}(1-q_k)} + \frac{1-\alpha_k}{1-q_k}\right) x_{k-1} \\ &= \frac{(\alpha_k - q_k)(1-\alpha_{k-1}^{-1})}{1-q_k} x_{k-2} + \left(\frac{(\alpha_{k-1}^{-1} - 1)(\alpha_k - q_k) + \alpha_k - q_k + 1 - \alpha_k}{1-q_k}\right) x_{k-1} \\ &= \frac{(\alpha_k - q_k)(1-\alpha_{k-1}^{-1})}{1-q_k} x_{k-2} + \left(\frac{(\alpha_{k-1}^{-1} - 1)(\alpha_k - q_k) g}{1-q_k} + 1\right) x_{k-1} \\ &= x_{k-1} + (1-q_k)^{-1}\left(\alpha_{k-1}^{-1} - 1\right)(\alpha_k - q_k)(x_{k-1} - x_{k-2}). \end{aligned}$$

■

3.1 preparations for the convergence rate proof

The following lemma summarize important results that gives a swift exposition for the proofs show up at the end for the convergence rate.

{lemma:cnvg-prep-part1} **Lemma 3.4 (convergence preparations part I)** *Let $(F, f, g, \mathcal{E}, \mu, L)$ satisfies Assumption 3.1. Suppose that*

(i) *The sequence $(y_k, v_k, x_k)_{k\geq 0}$ satisfies Definition 2.1 where T_B is defined on $F = f + g$.*
(ii) *The sequences $(\alpha_k)_{k\geq 0}, (\rho_k)_{k\geq 0}, (q_k)_{k\geq 0}$ satisfies the definition of relaxed momentum sequence.*

(iii) *We set the parameters* q_k *has* $q_k = \mu/B_k$*, with* $B_k > \mu$*, for all* $k \ge 0$*.*

Then, for all $\bar{x} \in \mathbb{R}^n$*,* $k \ge 0$*:*

$$
\begin{aligned}
&\frac{B_k\alpha_k^2}{2}\|\bar{x} - v_k\|^2 - \frac{B_k(1-\alpha_k)\rho_{k-1}\alpha_{k-1}^2}{2}\|\bar{x} - v_{k-1}\|^2 \\
&= \frac{\alpha_k\mu}{2}\|\bar{x} - v_{k-1}\|^2 + \frac{q_k\mu}{2}\|y_k - v_{k-1}\| + \frac{q_k}{2\mu}\|\mathcal{G}_{B_k}(y_k)\|^2 \\
&\quad - \langle q_k(y_k - v_{k-1}) + \alpha_k(v_{k-1} - \bar{x}), \mathcal{G}_{B_k}(y_k)\rangle + \alpha_k\mu\langle v_{k-1} - \bar{x}, y_k - v_{k-1}\rangle .
\end{aligned}
$$

Proof. ■

{lemma:cnvg-prep-part2}

Lemma 3.5 (convergence preparations part II)

{lemma:cnvg-prep-part3}

Lemma 3.6 (convergence preparations part III)

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.