Error Bound Can Give Near Optimal Convergence Rate for Inexact Accelerated Proximal Gradient Method

Author 1 Name, Author 2 Name *

October 16, 2025

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

Abstract

This is still a draft. [5].

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

1 Introduction

Notations. Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$, we denote $g^\star$ to be the Fenchel conjugate. $I : \mathbb{R}^n \to \mathbb{R}^n$ denotes the identity operator. For a multivalued mapping $T : \mathbb{R}^n \to 2^{\mathbb{R}^n}$, $\operatorname{gra} T$ denotes the graph of the operator, defined as $\{(x, y) \in \mathbb{R}^n \times \mathbb{R}^n : y \in Tx\}$.

1.1 Epsilon subgradient and inexact proximal point

{def:esp-subgrad}

Definition 1.1 (ϵ-subgradient) *Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper, lsc. Let $\epsilon \geq 0$. Then the ϵ-subgradient of g at some $\bar{x} \in \operatorname{dom} g$ is given by:*

$$\partial g_\epsilon(\bar{x}) := \{v \in \mathbb{R}^n \mid \langle v, x - \bar{x}\rangle \leq g(x) - g(\bar{x}) + \epsilon \ \forall x \in \mathbb{R}^n\}.$$

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

When $\bar{x} \notin \operatorname{dom} g$, it has $\partial g_\epsilon(\bar{x}) = \emptyset$.

Remark 1.2 $\partial_\epsilon g$ is a multivalued operator and, it's not monotone, unless $\epsilon = 0$, which makes it equivalent to Fenchel subgradient ∂g.

If we assume lsc, proper and convex g, we will now introduce results in the literatures that we will use.

{fact:esp-fenchel-ineq}

Fact 1.3 (ϵ-Fenchel inequality) *Let $\epsilon \geq 0$, then:*

$$x^* \in \partial_\epsilon f(\bar{x}) \iff f^\star(x^*) + f(\bar{x}) \leq \langle x^*, \bar{x} \rangle + \epsilon \implies \bar{x} \in \partial_\epsilon f^\star(x^*).$$

They are all equivalent if $f^{\star\star}(\bar{x}) = f(\bar{x})$.

Remark 1.4 The above fact is taken from Zalinascu [4, Theorem 2.4.2].

{def:inxt-pp}

We will now define inexact proximal point based on ϵ-subgradient

Definition 1.5 (inexact proximal point) *For all $x \in \mathbb{R}^n, \epsilon \geq 0, \lambda > 0$, $\tilde{x}$ is an inexact evaluation of proximal point at x, if and only if it satisfies:*

$$\lambda^{-1}(x - \tilde{x}) \in \partial_\epsilon g(\tilde{x}).$$

We denote it by $\tilde{x} \approx_\epsilon \operatorname{prox}_{\lambda g}(x)$.

{fact:resv-identity}

Remark 1.6 This definition is nothing new, for example see Villa et al. [3, Definition 2.1]

Fact 1.7 (the resolvant identity) *Let $T : \mathbb{R}^n \to 2^{\mathbb{R}^n}$, then it has:*

$$(I + T)^{-1} = (I - (I + T^{-1})^{-1}).$$

{thm:inxt-moreau-decomp}

Theorem 1.8 (inexact Moreau decomposition) *Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a closed, convex and proper function. It has the equivalence*

$$\tilde{y} \approx_\epsilon \operatorname{prox}_{\lambda^{-1} g^\star}(\lambda^{-1} y) \iff y - \lambda \tilde{y} \approx_\epsilon \operatorname{prox}_{\lambda g}(y).$$

Proof. Consider $\tilde{y} \approx_\epsilon \text{prox}_{\lambda^{-1}g^\star}(\lambda^{-1}y)$, then it has:

$$\begin{aligned}
&\tilde{y} \in (I + \lambda^{-1}\partial_\epsilon g^\star)^{-1}(\lambda^{-1}y) \\
\iff &(\lambda^{-1}y, \tilde{y}) \in \text{gra}(I + \lambda^{-1}\partial_\epsilon g^\star)^{-1} \\
\underset{(1)}{\iff} &(\lambda^{-1}y, \tilde{y}) \in \text{gra}(I - (I + \partial_\epsilon g \circ (\lambda I))^{-1}) \\
\iff &(\lambda^{-1}y, \lambda^{-1}y - \tilde{y}) \in \text{gra}(I + \partial_\epsilon g \circ (\lambda I))^{-1} \\
\iff &(\lambda^{-1}y - \tilde{y}, \lambda^{-1}y) \in \text{gra}(I + \partial_\epsilon g \circ (\lambda I)) \\
\iff &(y - \lambda\tilde{y}, \lambda^{-1}y) \in \text{gra}(\lambda^{-1}I + \partial_\epsilon g) \\
\iff &(y - \lambda\tilde{y}, y) \in \text{gra}(I + \lambda\partial_\epsilon g) \\
\iff &y - \lambda\tilde{y} \in (I + \lambda\partial_\epsilon g)^{-1}y \\
\iff &y - \lambda\tilde{y} \approx_\epsilon \text{prox}_{\lambda g}(y).
\end{aligned}$$

At (1) we can use Fact 1.7, and it has $(\lambda^{-1}\partial_\epsilon g^\star)^{-1} = \partial_\epsilon g \circ (\lambda I)$ by Fact 1.3 and the assumption that g is closed, convex and proper. ■

1.2 Inexact proximal gradient inequality

{ass:for-inxt-pg-ineq}

Assumption 1.9 (for inexact proximal gradient) The assumption is about (F, f, g, L). We assume that

(i) $f : \mathbb{R}^n \to \mathbb{R}$ is a convex, L Lipschitz smooth function (i.e: ∇f is L a Lipschitz continuous mapping).
(ii) $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a convex, proper, and lsc function which we do not have its exact proximal operator.
(iii) The over all objective is $F = f + g$.

No, we develop the theory based on the use of epsilon subgradient as in Definition 1.1. Let $\rho > 0$, the exact proximal gradient operator defined for (f, g, L) satisfying Assumption 1.9 has

$$\begin{aligned}
T_\rho(x) &= \underset{z\in\mathbb{R}^n}{\text{argmin}} \left\{ g(z) + \langle \nabla f(x), z\rangle + \frac{\rho}{2}\|z - x\|^2 \right\} \\
&= \text{prox}_{\rho^{-1}g}\left(x - \rho^{-1}\nabla f(x)\right).
\end{aligned}$$

{def:inxt-pg}

The following definition extends the proximal gradient operator to the inexact case using the concept of ϵ-subgradient as given by Definition 1.1.

Definition 1.10 (inexact proximal gradient) *Let (f, g, L) satisfies Assumption 1.9. Let $\epsilon \geq 0, \rho > 0$. Then, $\tilde{x} \approx_\epsilon T_\rho(x)$ is an inexact proximal gradient if it satisfies variational*

inequality:

$$\mathbf{0} \in \nabla f(x) + \rho(x - \tilde{x}) + \partial_\epsilon g(\tilde{x}).$$

Remark 1.11 We assumed that we can get exact evaluation of ∇f at any points $x \in \mathbb{R}^n$.

{lemma:other-repr-inxt-pg}

Lemma 1.12 (other representations of inexact proximal gradient) *Let (f, g, L) satisfies Assumption 1.9, $\epsilon \geq 0, \rho > 0$, then for all $\tilde{x} \approx_\epsilon T_\rho(x)$, it has the following equivalent representations:*

$$\begin{aligned} & (x - \rho^{-1}\nabla f(x)) - \tilde{x} \in \rho^{-1}\partial_\epsilon g(\tilde{x}) \\ \iff & \tilde{x} \in (I + \rho^{-1}\partial_\epsilon g(\tilde{x}))^{-1}(x - \rho^{-1}\nabla f(x)) \\ \iff & x \approx_\epsilon \operatorname{prox}_{\rho^{-1}g}\left(x - \rho^{-1}\nabla f(x)\right) \end{aligned}$$

Proof. It's direct. ■

{thm:inxt-pg-ineq}

Theorem 1.13 (inexact over-regularized proximal gradient inequality) *Let (f, g, L) satisfies Assumption 1.9, $\epsilon \geq 0, B \geq 0, \rho > 0$. Consider $\tilde{x} \approx_\epsilon T_{B+\rho}(x)$. Denote $F = f + g$. If in addition, $\tilde{x}, B$ satisfies the line search condition $D_f(\tilde{x}, x) \leq B/2\|x - \tilde{x}\|^2$, then it has $\forall z \in \mathbb{R}^n$:*

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

Proof. By Definition 1.10 write the variational inequality that describes $\tilde{x} \approx_\epsilon T_B(x)$, and the definition of epsilon subgradient (Definition 1.1) it has for all $z \in \mathbb{R}^n$:

$$\begin{aligned} -\epsilon &\leq g(z) - g(\tilde{x}) - \langle (B+\rho)(\tilde{x} - x) - \nabla f(x), z - \tilde{x}\rangle \\ &= g(z) - g(\tilde{x}) - (B+\rho)\langle \tilde{x} - x, z - \tilde{x}\rangle + \langle \nabla f(x), z - \tilde{x}\rangle \\ &\underset{(1)}{\leq} g(z) + f(z) - g(\tilde{x}) - f(\tilde{x}) - (B+\rho)\langle \tilde{x} - x, z - \tilde{x}\rangle - D_f(z, x) + D_f(\tilde{x}, x) \\ &\underset{(2)}{\leq} F(z) - F(\tilde{x}) - (B+\rho)\langle \tilde{x} - x, z - \tilde{x}\rangle + \frac{B}{2}\|\tilde{x} - x\|^2 \\ &= F(z) - F(\tilde{x}) + \frac{B+\rho}{2}\left(\|x - z\|^2 - \|\tilde{x} - x\|^2 - \|z - \tilde{x}\|^2\right) + \frac{B}{2}\|\tilde{x} - x\|^2 \\ &= F(z) - F(\tilde{x}) + \frac{B+\rho}{2}\|x - z\|^2 - \frac{B+\rho}{2}\|z - \tilde{x}\|^2 - \frac{\rho}{2}\|\tilde{x} - x\|^2. \end{aligned}$$

At (1), we used considered the following:

$$\begin{aligned} \langle \nabla f(x), z - x\rangle &= \langle \nabla f(x), z - x + x - \tilde{x}\rangle \\ &= \langle \nabla f(x), z - x\rangle + \langle \nabla f(x), x - \tilde{x}\rangle \\ &= -D_f(z, x) + f(z) - f(x) + D_f(\tilde{x}, x) - f(\tilde{x}) + f(x) \\ &= -D_f(z, x) + f(z) + D_f(\tilde{x}, x) - f(\tilde{x}). \end{aligned}$$

At (2), we used the fact that f is convex hence $-D_f(z,x) \leq 0$ always, and in the statement hypothesis we assumed that B has $D_f(\tilde{x},x) \leq B/2\|\tilde{x}-x\|^2$. We also used $F = f+g$. ■

Remark 1.14 When $\epsilon = 0, \rho = 0$, this reduces to proximal gradient inequality in the exact case. In this inequality, observe that the parameter ϵ controls the inexactness of the proximal gradient evaluation. More specifically, ϵ_k controls the absolute perturbations of the proximal gradient inequality compared to its exact counterpart. ρ on the other hand, it is the over-relaxation of proximal gradient operator, and it compensates the perturbations caused by ϵ relative to the term $\|\tilde{x}-x\|^2$.

1.3 Optimizing the inexact proximal point problem

{sec:optz-inxt-pp-problem}

In this section we will present the optimization problem that obtains a $\tilde{x}$ such that $\tilde{x} \approx_\epsilon \operatorname{prox}_{\lambda g}(z)$. Eventually we want to evaluate $T_\rho(x)$ of some $F = f+g$ inexactly using Lemma 1.12. To do that one would need to evaluate $\operatorname{prox}_{\rho^{-1}g}$ inexactly which is defined in Definition 1.5.

Most of these results that will follow are from the literature. To start, we must assume the following about a function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$, with g closed, convex and proper.

{ass:for-inxt-prox}

Assumption 1.15 (linear composite of convex nonsmooth function)
This assumption is about (g, ω, A). Let $m \in \mathbb{N}, n \in \mathbb{R}^n$, we assume that

(i) $A \in \mathbb{R}^{m\times n}$ is a matrix.
(ii) $\omega : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a closed and convex function such that it admits proximal operator $\operatorname{prox}_{\lambda\omega}$ and, its conjugate $\omega^\star$ is known.
(iii) $g := \omega(Ax)$ such that $\operatorname{rng} A \cap \operatorname{ri}\operatorname{dom} g \neq \emptyset$.

Now, we are ready to discuss how to choose $\tilde{x} \approx_\epsilon \operatorname{prox}_{\lambda g}(x)$. Fix $y \in \mathbb{R}^n, \lambda > 0$, we are ultimately interested in minimizing:

{eqn:primal-pp}

$$\Phi_\lambda(u) := \omega(Au) + \frac{1}{2\lambda}\|u-y\|^2 \tag{1.1}$$

This problem admits dual objective in $\mathbb{R}^m$:

{eqn:dual-pp}

$$\Psi_\lambda(v) := \frac{1}{2\lambda}\|\lambda A^\top v - y\|^2 + \omega^\star(v) - \frac{1}{2\lambda}\|y\|^2. \tag{1.2}$$

We define the duality gap

{eqn:duality-gap-pp}

$$\mathbf{G}_\lambda(u,v) := \Phi_\lambda(u) + \Psi_\lambda(v). \tag{1.3}$$

If strong duality holds, it exists $(\hat{u}, \hat{v})$ such that we have the following:

$$\mathbf{G}_\lambda(\hat{u}, \hat{v}) = 0 = \min_u \Phi_\lambda(u) + \min_v \Psi_\lambda(v)$$

{thm:primal-dual-trans}

The following theorem quantifies a sufficient conditions for $\tilde{x} \approx_\epsilon \operatorname{prox}_{\lambda g}(x)$. The theorem below is from [3, Proposition 2.2].

Theorem 1.16 (primal translate to dual [3, Proposition 2.2]) *Let (g, ω, A) satisfies assumption 1.15, $\epsilon \geq 0$, then*

$$\left(\forall z \approx_\epsilon \operatorname{prox}_{\lambda g}(y)\right)\left(\exists v \in \operatorname{dom} \omega^\star\right) : z = y - \lambda A^\top v.$$

This theorem that follows is from Villa et al. [3, Proposition 2.3].

{thm:dlty-gap-inxt-pp}

Theorem 1.17 (duality gap of inexact proximal problem [3, Proposition 2.3]) *Let (g, ω, A) satisfies Assumption 1.15, for all $\epsilon \geq 0$, $v \in \mathbb{R}^n$ consider the following conditions:*

(i) $\mathbf{G}_\lambda(y - \lambda A^\top v, v) \leq \epsilon$.
(ii) $A^\top v \approx_\epsilon \operatorname{prox}_{\lambda^{-1} g^\star}(\lambda^{-1} y)$.
(iii) $y - \lambda A^\top v \approx_\epsilon \operatorname{prox}_{\lambda g}(y)$.

They have $(a) \implies (b) \iff (c)$. If in addition $\omega^\star(v) = g^\star(A^\top v)$, then all three conditions are equivalent.

Proof. The proof of $(a) \implies (b)$, and the case when $(a) \iff (b)$, we refer readers to Villa et al. [3, Proposition 2.3], and to show $(b) \iff (c)$ use Theorem 1.8. ■

{fact:minimizing-dual-pp}

The following fact from the literature indicates that it's sufficient to minimize the dual problem Ψ_λ to obtain an element of the inexact proximal point operator. The following fact is Proposition is from Villa et al. [3, Theorem 5.1].

Fact 1.18 (minimizing dual of the proximal problem [3, Theorem 5.1]) *Let $\bar{v}$ be a solution of Ψ_λ. Suppose that $(v_n)_{n\geq 0}$ is a minimizing sequence for Ψ_λ. Let $z_n = y - \lambda A^\top v_n$, and $\bar{z} = y - \lambda A^\top \bar{v}$. If in addition, Φ_λ is L_1 Lipschitz continuous, then it has for all $k \geq 0$ the inequality:*

$$\Phi_\lambda(z_n) - \Phi_\lambda(\bar{z}) \leq L_1 \|z_n - \bar{z}\| \leq L_1 \sqrt{2\lambda} (\Psi_\lambda(v_n) - \Psi_\lambda(\bar{v}))^{1/2}.$$

We remark that the above fact translates any algorithm that optimizes the function value of the dual problem Ψ_λ into optimizing duality gap $\mathbf{G}(z_n, v_n)$. For this reason, the number of iterations of the inner loop required to achieve $\mathbf{G}(z_n, v_n) < \epsilon$ for a given ϵ is related to the convergence rate of the algorithms used to optimize $\Psi_\lambda(v_n)$. With the theorem derived above, and using Theorem 1.17 it implies that any algorithm which can optimize function value Ψ_λ will produce iterates sufficient to achieve $\approx_\epsilon \operatorname{prox}_{\lambda g}(y)$.

1.4 Literature reviews

1.5 Our contributions

2 The inexact accelerated proximal gradient with controlled errors

In this section, we present an accelerated algorithm with controlled error using Definition 1.10, and show that it can have a convergence rate under certain error conditions.

{def:inxt-apg}

Definition 2.1 (our inexact accelerated proximal gradient) *Suppose that (F, f, g, L) and, sequences $(\alpha_k, B_k, \rho_k, \epsilon_k)_{k\geq 0}$ satisfies the following*

(i) *$(\alpha_k)_{k\geq 0}$ is a sequence such that $\alpha \in (0, 1]$ for all $k \geq 0$.*
(ii) *$(B_k)_{k\geq 0}$ has $B_k > 0\ \forall k$, it characterizes any potential line search, back tracking routine.*
(iii) *$(\rho_k)_{k\geq 0}$ be a sequence such that $\rho_k \geq 0$, characterizing the over-relaxation of the proximal gradient operator.*
(iv) *$(\epsilon_k)_{k\geq 0}$ has $\epsilon_k > 0$ for all $k \geq 0$, it characterizes the errors of inexact proximal evaluation.*
(v) *(f, g, L) satisfies Assumption 1.9, and let $F = f + g$.*

Denote $L_k = B_k + \rho_k$ for short. Given any initial condition $v_{-1}, x_{-1} \in \mathbb{R}^n$, the algorithm generates the sequences $(y_k, x_k, v_k)_{k\geq 0}$ such that they satisfy for all $k \geq 0$:

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

{lemma:inxt-apg-cnvg-prep1}

Lemma 2.2 (inexact accelerated proximal gradient preparation stage I) *Let (F, f, g, L), and $(\alpha_k, B_k, \rho_k, \epsilon_k)_{k\geq 0}$, be given by Definition 2.1. Denote $L_k = B_k + \rho_k$. Then, for any $\bar{x} \in \mathbb{R}^n$, the sequences $(y_k, x_k, v_k)_{k\geq 0}$ generated satisfy for all $k \geq 1$ the inequality:*

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

When, $k = 1$ it instead has:

$$\begin{aligned}&\frac{\rho_0}{2}\|x_0 - y_0\|^2 - \epsilon_0\\ &\leq (1-\alpha_0)(F(x_{-1}) - F(\bar{x})) + F(\bar{x}) - F(x_0) + \frac{\alpha_0^2 L_0}{2}\|\bar{x} - v_{-1}\|^2 - \frac{\alpha_0^2 L_0}{2}\|\bar{x} - v_0\|^2.\end{aligned}$$

Proof. Two intermediate results are in order before we can prove the inequality. Define $z_k := \alpha_k \bar{x} + (1-\alpha_k)x_{k-1}$ for short. It has for all $k \geq 1$ the equality:

{eqn:inxt-apg-cnvg-prep1-a}

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

It also has for all $k \geq 1$ the equality:

{eqn:inxt-apg-cnvg-prep1-b}

$$\begin{aligned}z_k - y_k &= \alpha_k \bar{x} + (1-\alpha_k)x_{k-1} - y_k\\ &= \alpha_k \bar{x} - \alpha_k v_{k-1}.\end{aligned} \tag{b}$$

Let's denote $L_k = B_k + \rho_k$ for short. Recall that (f, g, L) satisfies Assumption 1.9, if we choose $x = y_k$ so $\tilde{x} = x_k \approx_{\epsilon_k} T_{L_k}(y_k)$, and set $z = z_k, \epsilon = \epsilon_k$ then Theorem 1.13 has:

$$\begin{aligned}&\frac{\rho_k}{2}\|x_k - y_k\|^2 - \epsilon_k\\ &\leq F(z_k) - F(x_k) + \frac{L_k}{2}\|y_k - z_k\|^2 - \frac{L_k}{2}\|z_k - x_k\|^2\\ &\underset{(1)}{\leq} \alpha_k F(\bar{x}) + (1-\alpha_k)F(x_{k-1}) - F(x_k) + \frac{L_k}{2}\|y_k - z_k\|^2 - \frac{L_k}{2}\|z_k - x_k\|^2\\ &\underset{(2)}{=} (1-\alpha_k)(F(x_{k-1}) - F(\bar{x})) + F(\bar{x}) - F(x_k) + \frac{\alpha_k^2 L_k}{2}\|\bar{x} - v_{k-1}\|^2 - \frac{\alpha_k^2 L_k}{2}\|\bar{x} - v_k\|^2\\ &\leq (1-\alpha_k)(F(x_{k-1}) - F(\bar{x})) + F(\bar{x}) - F(x_k)\\ &+ \max\left(1-\alpha_k, \frac{\alpha_k^2 L_k}{\alpha_{k-1}^2 L_{k-1}}\right)\frac{\alpha_{k-1}^2 L_{k-1}}{2}\|\bar{x} - v_{k-1}\|^2 - \frac{\alpha_k^2 L_k}{2}\|\bar{x} - v_k\|^2.\end{aligned}$$

At (1) we used the fact that $F = f + g$ hence F is convex. At (2) we used (a), (b). Finally, if $k = 0$, then take the RHS of $\underset{(1)}{=}$ then:

$$\begin{aligned}&\frac{\rho_0}{2}\|x_0 - y_0\|^2 - \epsilon_0\\ &\leq (1-\alpha_0)(F(x_{-1}) - F(\bar{x})) + F(\bar{x}) - F(x_0) + \frac{\alpha_0^2 L_0}{2}\|\bar{x} - v_{-1}\|^2 - \frac{\alpha_0^2 L_0}{2}\|\bar{x} - v_0\|^2.\end{aligned}$$

■

The following assumption encapsulate assumptions on the errors such that a near optimal convergence rate is still attainable by an algorithm that satisfies Definition 2.1.

{ass:valid-err-schedule}

Assumption 2.3 (valid error schedule) The following assumption is about an algorithm satisfying Definition 2.1, its parameters $(\alpha_k, B_k, \rho_k, \epsilon_k)_{k\geq 0}$ in relation to its iterates $(y_k, x_k, v_k)_{k\geq 0}$ and, some additional parameters $(\beta_k)_{k\geq 0}, \mathcal{E}_0$ and p. Let

(i) $(\alpha_k, B_k, \rho_k, \epsilon_k)_{k\geq 0}, (F, f, g, L)$ and $(y_k, x_k, v_k)_{k\geq 0}$ be given by Definition 2.1.
(ii) $\mathcal{E}_0 \geq 0$ be arbitrary;
(iii) the sequence $(\beta_k)_{k\geq 0}$ be defined as $\beta_k := \prod_{i=1}^{k} \max\left(1-\alpha_i, \frac{\alpha_i^2 L_i}{\alpha_{i-1}^2 L_{i-1}}\right)$ for all $k \geq 1$, with the base case being $\beta_0 = 1$;
(iv) $p \geq 1$ is some constant which will bound the error ϵ_k relative to $\rho_k\|x_k - y_k\|^2, \beta_k$ and, k.

In addition, we assume that the error parameter $\epsilon_k \geq 0$ and over-relaxation parameter ρ_k, iterates x_k, y_k and β_k together satisfies for all $k \geq 0$ the relations:

$$\frac{-\mathcal{E}_0\beta_k}{k^p} \leq \frac{\rho_k}{2}\|x_k - y_k\|^2 - \epsilon_k.$$

The following proposition is a prototype of the convergence rate together with the error schedule that delivers convergence of algorithms satisfying Definition 2.1.

{prop:inxt-apg-cnvg-generic}

Proposition 2.4 (generic convergence rate under valid error schedule) *Let (F, f, g, L), $(\alpha_k, B_k, \rho_k, \epsilon_k)_{k\geq 0}$, $(\beta_k)_{k\geq 0}, \mathcal{E}_0, p$ as assumed in Assumption 2.3. Fix any $\bar{x} \in \mathbb{R}^n$ for all $k \geq 0$ and assume that $\alpha_0 = 1$. Then for the iterates generated $(y_k, x_k, v_k)_{k\geq 0}$ by the algorithm, for all $k \geq 0$ they will satisfy:*

$$F(x_k) - F(\bar{x}) + \frac{\alpha_k^2 L_k}{2}\|\bar{x} - v_k\|^2 \leq \beta_k\left(\frac{L_0}{2}\|\bar{x} - v_{-1}\|^2 + \mathcal{E}_0\sum_{n=1}^{k}\frac{1}{n^p}\right).$$

Proof. Consider results from Lemma 2.2 has $\forall k \geq 1$:

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

For notation brevity, we introduce β_k, Λ_k:

$$\begin{aligned}
\beta_0 &= 1, \\
\beta_k &:= \prod_{i=1}^{k} \max\left(1-\alpha_i, \frac{\alpha_i^2 L_i}{\alpha_{i-1}^2 L_{i-1}}\right), \\
\Lambda_k &:= -F(\bar{x}) + F(x_k) + \frac{\alpha_k^2 L_k}{2}\|\bar{x} - v_k\|^2.
\end{aligned}$$

Now, suppose that in addition there is a non-negative sequence $(\mathcal{E}_k)_{k\geq 0}$ such that

(i) For all $k \geq 0$, it has $\frac{-\mathcal{E}_k}{k^p} \leq \frac{\rho_k}{2}\|x_k - y_k\|^2 - \epsilon_k$ where $p \geq 1$,
(ii) For all $k \geq 1$, it has $\mathcal{E}_k = \frac{\beta_k}{\beta_{k-1}}\mathcal{E}_{k-1}$, with $\mathcal{E}_0 \geq 0$.

These conditions are equivalent to the assumption that $\frac{-\mathcal{E}_0\beta_k}{k^p} \leq \frac{\rho_k}{2}\|x_k - y_k\|^2 - \epsilon_k$ (which was stated in Assumption 2.3). One can show that by unrolling recurrence on $\mathcal{E}_k$. Then (2.1) implies $\forall k \geq 1$:

{ineq:inxt-apg-cnvg-generic-pitem-1}

$$\frac{-\mathcal{E}_k}{k^p} \leq \frac{\beta_k}{\beta_{k-1}}\Lambda_{k-1} - \Lambda_k \iff \Lambda_k \leq \frac{\beta_k}{\beta_{k-1}}\Lambda_{k-1} + \frac{\mathcal{E}_k}{k^p}. \tag{2.1}$$

Now, we show the convergence of Λ_k, using the relations of $\mathcal{E}_k, \Lambda_k, \beta_k$ above.

$$\begin{aligned}
\Lambda_k &\leq \frac{\beta_k}{\beta_{k-1}}\Lambda_{k-1} + \frac{\mathcal{E}_k}{k^p} \\
&\leq \frac{\beta_k}{\beta_{k-1}}\Lambda_{k-1} + \frac{\beta_k}{\beta_{k-1}}\frac{\mathcal{E}_{k-1}}{k^p} \\
&= \frac{\beta_k}{\beta_{k-1}}\left(\Lambda_{k-1} + \frac{\mathcal{E}_{k-1}}{k^p}\right) \\
&\leq \frac{\beta_k}{\beta_{k-1}}\left(\frac{\beta_{k-1}}{\beta_{k-2}}\Lambda_{k-2} + \frac{\mathcal{E}_{k-1}}{(k-1)^p} + \frac{\mathcal{E}_{k-1}}{k^p}\right) \\
&= \frac{\beta_k}{\beta_{k-2}}\left(\Lambda_{k-2} + \frac{\mathcal{E}_{k-2}}{(k-1)^p} + \frac{\mathcal{E}_{k-2}}{k^p}\right) \\
&\cdots \\
&\leq \frac{\beta_k}{\beta_1}\left(\Lambda_1 + \mathcal{E}_1\sum_{n=2}^{k}\frac{1}{n^p}\right) \\
&\leq \frac{\beta_k}{\beta_1}\left(\frac{\beta_1}{\beta_0}\Lambda_0 + \mathcal{E}_1\sum_{n=1}^{k}\frac{1}{n^p}\right) \\
&= \frac{\beta_k}{\beta_0}\left(\Lambda_0 + \mathcal{E}_0\sum_{n=1}^{k}\frac{1}{n^p}\right).
\end{aligned}$$

Therefore, it points to the following inequality:

$$F(x_k) - F(\bar{x}) + \frac{\alpha_k^2 L_k}{2}\|\bar{x} - v_k\|^2 \\ \leq \beta_k \left(F(x_0) - F(\bar{x}) + \frac{\alpha_0^2 L_0}{2}\|\bar{x} - v_0\|^2 + \mathcal{E}_0 \sum_{n=1}^{k} \frac{1}{n^p} \right).$$

Finally, when $\alpha_0 = 1$, then the results from 2.2 with $k = 0$ simplifies the above inequality and give:

$$F(x_k) - F(\bar{x}) + \frac{\alpha_k^2 L_k}{2}\|\bar{x} - v_k\|^2 \leq \beta_k \left(\frac{L_0}{2}\|\bar{x} - v_{-1}\|^2 + \mathcal{E}_0 \sum_{n=1}^{k} \frac{1}{n^p} \right).$$

■

Now, it only remains to determine the sequence α_k to derive a type of convergence rate for the algorithm because from the above theorem, we have the convergence rate β_k and, the error parameters ϵ_k, ρ_k both controlled by the sequence $(\alpha_k)_{k\geq 0}$.

2.1 Convergence results of the outer loop

This section will give specific instances of the error control sequence $(\epsilon_k)_{k\geq 0}, (\rho_k)_{k\geq 0}$ and, momentum sequence $(\alpha_k)_{k\geq 0}$ such that an optimal convergence rate of $\mathcal{O}(1/k^2)$ can be achieved.

{ass:opt-mmntm-seq}

Assumption 2.5 (the optimal momentum sequence)
Keeping everything assumed in Assumption 2.3 about $(\alpha_k, B_k, \rho_k, \epsilon_k)_{k\geq 0}$, (F, f, g, L), $(y_k, x_k, v_k)_{k\geq 0}$, $(\beta_k)_{k\geq 0}$, $\mathcal{E}_0$ and p, we assume in addition that the sequence $(\alpha_k)_{k\geq 0}$ satisfies for all $k \geq 0$ the equality: $(1 - \alpha_k) = \alpha_k^2 L_k \alpha_{k-1}^{-2} L_{k-1}^{-1}$.

{lemma:opt-mmntm-seq}

Lemma 2.6 (the optimal momentum sequence is indeed valid and optimal)
Let $(\alpha_k)_{k\geq 0}, (\beta_k)_{k\geq 0}$ be given by Assumption 2.5. If we choose $\alpha_0 \in (0, 1]$ then it satisfies $\alpha_k \in (0, 1)$ $\forall k \geq 1$, and it's given by:

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

In addition, the sequence $(\beta_k)_{k\geq 0}$ has

$$\left(1 + \alpha_0\sqrt{L_0}\sum_{i=1}^{k}\sqrt{L_i^{-1}} \right)^{-2} \leq \beta_k \leq \left(1 + \frac{\alpha_0\sqrt{L_0}}{2}\sum_{i=1}^{k}\sqrt{L_i^{-1}} \right)^{-2}.$$

Proof. Firstly, we will show that for all $\alpha_0 \in (0,1]$, and $\alpha_k \in (0,1)$ $\forall k \geq 1$. We will prove using induction. Assume inductively that $\alpha_k \in (0,1]$. We can solve for α_k using the recursive equality from Assumption 2.5 which always has one of the strictly positive root given by:

$$\begin{aligned}
\alpha_k &= \frac{L_{k-1}}{2L_k}\left(-\alpha_{k-1}^2 + \left(\alpha_{k-1}^4 + 4\alpha_{k-1}\frac{L_k}{L_{k-1}}\right)^{1/2}\right) \\
&= \frac{\alpha_{k-1}^2 L_{k-1}}{2L_k}\left(-1 + \left(1 + \frac{4L_k}{L_{k-1}\alpha_{k-1}^2}\right)^{1/2}\right) \\
&\underset{(1)}{<} \frac{\alpha_{k-1}^2 L_{k-1}}{2L_k}\left(-1 + 1 + 2\alpha_{k-1}^{-2}\frac{L_k}{L_{k-1}}\right) \\
&= 1
\end{aligned}$$

At (1) we completed a square inside the radical, and use the assumption that $\alpha_k > 0$ and, $L_k > 0, L_{k-1} > 0$ because we had $B_k > 0$:

$$\begin{aligned}
1 + 4L_k L_{k-1}^{-1}\alpha_{k-1}^{-2} &= 1 + 4L_k L_{k-1}^{-1}\alpha_{k-1}^{-2} + 2L_k^2 L_{k-1}^{-2}\alpha_{k-1}^{-4} - 2L_k^2 L_{k-1}^{-2}\alpha_{k-1}^{-4} \\
&= (1 + 2L_k L_{k-1}^{-1}\alpha_{k-1}^{-2})^2 - 2\alpha_{k-1}^{-4}L_k^2 L_{k-1}^{-2} \\
&< \left(1 + 2L_k L_{k-1}^{-1}\alpha_{k-1}^{-2}\right)^2.
\end{aligned}$$

To see that $\alpha_k > 0$, recall the same fact that $L_k > 0$, and use the inductive hypothesis that $\alpha_{k-1} \in (0,1]$ then $\alpha_k > 0$ because:

$$\begin{aligned}
\alpha_k &= \frac{L_{k-1}}{2L_k}\left(-\alpha_{k-1}^2 + \left(\alpha_{k-1}^4 + 4\alpha_{k-1}\frac{L_k}{L_{k-1}}\right)^{1/2}\right) \\
&\geq \frac{L_{k-1}}{2L_k}\left(-\alpha_{k-1}^2 + \left(\alpha_{k-1}^4\right)^{1/2}\right) \\
&> 0.
\end{aligned}$$

Therefore, inductively it holds that $\alpha_k \in (0,1)$ too.

Now, we focus on $(\beta_k)_{k\geq 0}$. From the assumption that $(\alpha_k)_{k\geq 0}$ has $(1-\alpha_k) = \alpha_k^2 L_k \alpha_{k-1}^{-2} L_{k-1}^{-1}$ for all $k \geq 0$ and, the definition of β_k, it yields the following equalities:

$$\beta_k = \prod_{i=1}^{k} \max\left(1 - \alpha_i, \frac{\alpha_i^2 L}{\alpha_{i-1}^2 L_{i-1}}\right) = \prod_{i=1}^{k}(1-\alpha_i) = \prod_{i=1}^{k}\frac{\alpha_i^2 L}{\alpha_0^2 L_0} = \frac{\alpha_k^2 L_k}{\alpha_0^2 L_0}.$$

With the above relation, and the definitions of the sequences $(\alpha_k)_{k\geq 0}, (\beta_k)_{k\geq 0}$ it satisfies for all $k \geq 1$ the properties:

(a) β_k is monotone decreasing and $\beta_k > 0$ for all $k \geq 0$ because $\beta_k = \prod_{i=1}^{k}(1-\alpha_i)$ and, $\alpha_k \in (0,1]$.

(b) It has the equalities $\beta_k/\beta_{k-1} = (1-\alpha_k) = \frac{\alpha_k^2 L_k}{\alpha_{k-1}^2 L_{k-1}}$ for all $k \geq 1$.

Using the above observations, we can show the chain of equalities $\alpha_k^2 = (1-\beta_k/\beta_{k-1})^2 = \beta_k L_0 \alpha_0^2 L_k^{-1}$ for all $k \geq 0$. This is true by first considering the relations $\prod_{i=1}^k (1-\alpha_i) = \beta_k$:

{eqn:opt-mmntm-seq-pitem1}

$$\begin{aligned} & (1-\alpha_k) = \beta_k/\beta_{k-1} \\ \iff & \alpha_k = 1 - \beta_k/\beta_{k-1} \\ \implies & \alpha_k^2 = (1-\beta_k/\beta_{k-1})^2. \end{aligned} \tag{2.2}$$

Next, the recursive relation of $(\alpha_k)_{k\geq 0}$ gives

{eqn:opt-mmntm-seq-pitem2}

$$\begin{aligned} \alpha_k^2 &= (1-\alpha_k)\alpha_{k-1}^2 L_{k-1} L_k^{-1} \\ &= (1-\alpha_k)\frac{\alpha_{k-1}^2 L_{k-1}}{\alpha_0^2 L_0}\frac{L_{k-1}\alpha_0^2 L_0}{L_k} \\ &= (\beta_k \beta_{k-1}^{-1})\beta_{k-1} L_0 \alpha_0^2 L_k^{-1} \\ &= \beta_k L_0 \alpha_0^2 L_k^{-1}. \end{aligned} \tag{2.3}$$

Combining (2.2), (2.3) and the fact that $\beta_k > 0\ \forall k \geq 0$, it would mean for all $i \geq 1$ it has:

$$\begin{aligned} L_0 \alpha_0^2 L_i^{-1} &= \beta_i^{-1}\left(1 - \frac{\beta_k}{\beta_{k-1}}\right)^2 \\ &= \beta_i \left(\beta_i^{-1} - \beta_{i-1}^{-1}\right)^2 \\ &= \beta_i \left(\beta_i^{-1/2} - \beta_{i-1}^{-1/2}\right)^2 \left(\beta_i^{-1/2} + \beta_{i-1}^{-1/2}\right)^2 \\ &= \left(\beta_i^{-1/2} - \beta_{i-1}^{-1/2}\right)^2 \left(1 + \beta_i^{1/2}\beta_{i-1}^{-1/2}\right)^2. \end{aligned}$$

Since β_i is monotone decreasing, it has $0 < \beta_i^{-1/2}\beta_{i-1}^{-1/2} \leq 1$, this gives:

$$\beta_i^{-1/2} - \beta_{i-1}^{-1/2} \leq \alpha_0 \sqrt{\frac{L_0}{L_i}} \leq 2\left(\beta_i^{-1/2} - \beta_{i-1}^{-1/2}\right).$$

Performing a telescoping sum, use the fact that $\beta_0 = 1$ will yield the desired results after some algebraic manipulations. ■

{prop:opt-cnvg-outr-loop}

Proposition 2.7 ($\mathcal{O}(1/k^2)$ optimal convergence rate of the outer loop) *Let (f, g, L), $(\alpha_k, B_k, \rho_k, \epsilon_k)_{k\geq 0}$, $(\beta_k)_{k\geq 0}, \mathcal{E}_0, p$ be given by Assumption 2.5. Assume in addition that*

(i) *there exists $\bar{x} \in \mathbb{R}^n$ that is a minimizer of $F = f + g$;*
(ii) *it has $p > 0$;*

(iii) *the sequence* $L_k := B_k + \rho_k$ *is bounded, and there exists an* $L_{\max}$ *such that for all* $k \geq 0$ *it has* $L_{\max} \geq \max_{k \geq i \geq 0} L_i$.

Then, $\alpha_k \in (0, 1]$ $\forall k \geq 0$ *hence it's always valid and, it has* $\forall k \geq 0$*:*

$$F(x_k) - F(\bar{x}) + \frac{\alpha_k^2 L_k}{2}\|\bar{x} - v_k\|^2 \leq \left(1 + \frac{k\alpha_0\sqrt{L_0}}{2\sqrt{L_{\max}}}\right)^{-2}\left(\frac{L_0}{2}\|\bar{x} - v_{-1}\|^2 + \mathcal{E}_0\sum_{n=1}^{k}\frac{1}{n^p}\right).$$

Since, $p > 0$ *the sum is convergent and hence the above inequality claims an overall convergence rate* $\mathcal{O}(1/k^2)$*.*

Proof. We use the results from 2.6 because the sequence has the same assumption, then using the fact that L_k is bounded then it gives:

$$\beta_k \leq \left(1 + \frac{\alpha_0\sqrt{L_0}}{2}\sum_{i=1}^{k}\sqrt{L_i^{-1}}\right)^{-2} \leq \left(1 + \frac{k\alpha_0\sqrt{L_0}}{2\sqrt{L_{\max}}}\right)^{-2}.$$

Then, apply Theorem 2.4 to obtain the desired results. ■

Remark 2.8 In this remark, we assert the fact that all assumptions made in the theorem can be satisfied on practice, and we will also bring attention to the current, and future roles played by each of the parameters used in the inexact algorithm.

Pay attention that α_k had been determined in the above theorem (as seemed in Lemma 2.6), and $(B_k)_{k \leq 0}$ is reserved for potential line search routine, the only parameter left undetermined in Definition 2.1 is the over-relaxation sequence $(\rho_k)_{k \geq 0}$, we only need the entire sequence to be bounded above. We give the freedom to the practitioners to choose $(\rho_k)_{k \geq 0}$. However, this sequence ρ_k is not useless because it relaxes the upper bound for ϵ_k, this has huge impact in the earlier stage (the first few iterations) of the algorithm $\|x_k - y_k\|$ is large. Of course, the parameter $\mathcal{E}_0$ is also free to choose.

Finally, observe that from the above proof, in case when $p = 1$, the convergence rate would be $\mathcal{O}(\log(k)/k^2)$.

In the next subsection, we will show that under the assumption of the above theorem, there exists an error sequence ϵ_k such that it can never approach 0 at an arbitrarily fast rate.

2.2 The fastest rate of which the error schedule can shrink

To have overall convergence claim of the algorithm, it's necessary to prevent the error schedule $(\epsilon_k)_{k \geq 0}$ from crashing into zero too quickly. Following Assumption 2.3, in this section,

we will provide the lower bound results for ϵ_k in Theorem 2.7 to show that in the worst case it cannot approach zero arbitrarily fast, if we choose the largest possible ϵ_k using β_k.

{lemma:err-schedule-lbnd}

Lemma 2.9 (error schedule lower bound)
Let, $(\alpha_k, B_k, \rho_k, \epsilon_k)_{k\geq 0}$, $(\beta_k)_{k\geq 0}, \mathcal{E}_0, p$ be given by Assumption 2.5, $L_k := \rho_k + B_k$. Let $(\epsilon_k)_{k\geq 0}$ be given by $\epsilon_k := \frac{\mathcal{E}_0\beta_k}{k^p} + \rho_k\frac{\|x_k - y_k\|^2}{2}$, then it will be a valid error sequence and so that it satisfies the assumption. If in addition, there exists $L_{\min}$ such that for all $k \geq 0$ such that it has $L_{\min} \leq \min_{1\leq i\leq k} L_i$ and, we assume $\mathcal{E} > 0$, then ϵ_k admits the non-trivial lower bound:

$$\epsilon_k \geq \frac{\mathcal{E}_0}{k^p}\left(1 + k\alpha_0\sqrt{L_0}\sqrt{L_{\min}^{-1}}\right)^{-2} \geq \mathcal{O}(k^{-2-p}).$$

Proof. Recall Assumption 2.3, the largest valid error schedule is $\epsilon_k = \frac{\mathcal{E}_0\beta_k}{k^p} + \rho_k\frac{\|x_k - y_k\|^2}{2}$. Then it has

$$\begin{aligned}
\epsilon_k &\geq \frac{\mathcal{E}_0\beta_k}{k^p} \\
&\underset{(1)}{\geq} \left(1 + \alpha_0\sqrt{L_0}\sum_{i=1}^{k}\sqrt{L_i^{-1}}\right)^{-2}\frac{\mathcal{E}_0\beta_k}{k^p} \\
&\underset{(2)}{\geq} \frac{\mathcal{E}_0\beta_k}{k^p}\left(1 + k\alpha_0\sqrt{L_0}\sqrt{L_{\min}^{-1}}\right)^{-2} \\
&\geq \mathcal{O}(k^{-2-p}).
\end{aligned}$$

At (1), we used Lemma 2.6. At (2), we used that $L_{\min} \leq L_i$ for all $i = 0, 1, 2, \ldots$. ∎

3 Linear convergence for the proximal problem in the inner loop

In this section, we continue the discussion from Section 1.3. As an important reminder, we will fix the vector $y \in \mathbb{R}^n$, which is in the inexact proximal problem as a constant in this entire section.

The inner loop of the algorithm is another algorithm that evaluates $x_k \approx_\epsilon T_{(B_k+\rho_k)}(y_k)$ for a given value of $\epsilon, B + \rho$ and at the point y_k. To accomplish, let $\lambda = (B_k + \rho_k)^{-1}$, the algorithm needs to resolve the following equivalent inexact proximal point problem:

$$x_k \approx_\epsilon \operatorname{prox}_{\lambda g}(y_k - \lambda\nabla f(y_k)).$$

Unfortunately recall that $g = \omega \circ A$ in the context of the outer loop hence it's impossible to directly evaluate the proximal operator of g and hence we optimize the function Φ_λ as given by (1.1).

In this section, we will show that there exists an algorithm generating the sequences z_n, v_n such that $\mathbf{G}_\lambda(z_n, v_n)$ converges linearly if Ψ_λ satisfies the error bound conditions. Using results available in the literature, we will characterize the exact scenarios of $\omega \circ A$ when this is possible to achieve. To start, the following assumption is the general error bound condition of a convex with additive composite structure.

{ass:pg-eb}

Assumption 3.1 (gradient mapping error bound) The following assumption is about (F, f, g, L, S, γ). Assume that

(i) (f, g, L) satisfies Assumption 1.9,
(ii) let $\tau > 0$ be the step size inverse, let T_τ be the proximal operator of $f + g$ as given by $T_\tau(x) := \operatorname{prox}_{\tau^{-1}g}(x - \tau^{-1}\nabla f(x))$,
(iii) $S = \operatorname*{argmin}_x f(x) + g(x) \neq \emptyset$,
(iv) the objective function is given by $F = f + g$.

Let the gradient mapping $\mathcal{G}_\tau$ be defined as: $\mathcal{G}_\tau(x) := \tau(x - T_\tau(x))$ In addition, assume that the optimization problem F satisfies the error bound condition if it has for all $\tau \geq L, x \in \mathbb{R}^n$ there exists $\gamma > 0$:

$$\|\mathcal{G}_\tau(x)\| \geq \gamma \operatorname{dist}(x|S).$$

{def:ista}

Definition 3.2 (proximal gradient method) *Suppose that (f, g, L) satisfies Assumption 1.9. Let $\tau \geq L$, and $x_0 \in \mathbb{R}^n$. Then an algorithm is a proximal gradient method if it generates iterates $(x_k)_{k\geq 0}$ such that they satisfy for all $k \geq 1$:*

$$x_{k+1} = \operatorname{prox}_{\tau^{-1}g}\left(x_k + \tau^{-1}\nabla f(x_k)\right).$$

3.1 Error bound and linear convergence of ISTA

The following theorem characterizes linear convergence of the proximal gradient method under gradient mapping error bound condition.

{thm:lin-cnvg-ista-eb}

Theorem 3.3 (linear convergence under gradient mapping error bound) *Assume that (F, f, g, L, S, γ) is given by Assumption 3.1. Under this assumption, the iterates $(x_k)_{k\geq 0}$ given by Definition 3.2 satisfies for all $k \geq 0, \bar{x} \in S$ and $\tau \geq L$ the inequality:*

$$F(x_{k+1}) - F(\bar{x}) \leq \left(1 - \frac{\gamma}{2\tau}\right)(F(x_k) - F(\bar{x})).$$

Hence, the algorithm generates $F(x_k) - F(\bar{x}) \leq \mathcal{O}((1 - \gamma/(2\tau))^k)$.

Proof. Two important immediate results will be presented first. Consider the proximal gradient inequality from Theorem 1.13, but with $\rho = 0, \epsilon = 0, B = \tau$, then for all x such that $\|\mathcal{G}_\tau(x)\| > 0$ it has for $\tilde{x} = T_\tau(x), z \in \mathbb{R}^n$ the inequality

$$\begin{aligned} F(\tilde{x}) - F(z) &\leq \frac{\tau}{2}\|x - z\|^2 - \frac{\tau}{2}\|z - \tilde{x}\|^2 \\ &= -\frac{\tau}{2}\|x - \tilde{x}\|^2 + \tau\langle x - z, x - \tilde{x}\rangle \\ &= -\frac{1}{2\tau}\|\mathcal{G}_\tau(x)\|^2 + \langle x - z, \mathcal{G}_\tau(x)\rangle \\ &\leq -\frac{1}{2\tau}\|\mathcal{G}_\tau(x)\|^2 + \|x - z\|\|\mathcal{G}_\tau(x)\| \\ &= \|\mathcal{G}_\tau(x)\|^2 \left(\frac{\|x - z\|}{\|\mathcal{G}_\tau(x)\|} - \frac{1}{2\tau}\right). \end{aligned}$$

Now, for all $z = \bar{x} \in S$, from Assumption 3.4 $\exists\gamma > 0$ such that:

$$\frac{\|x - z\|}{\|\mathcal{G}_\tau(x)\|} \leq \frac{\|x - z\|}{\gamma \operatorname{dist}(x|S)} \leq \frac{1}{\gamma}.$$

Hence, for all $\bar{x} \in S$ it has

{ineq:lin-cnvg-ista-eb-pitem1}
$$0 \leq F(\tilde{x}) - F(\bar{x}) \leq \|\mathcal{G}_\tau(x)\|^2 \left(\frac{1}{\gamma} - \frac{1}{2\tau}\right). \tag{3.1}$$

Obviously it has $\gamma^{-1} - (1/2)\tau^{-1} > 0$. When $z = x$, we have the inequality:

{ineq:lin-cnvg-ista-eb-pitem2}
$$F(\tilde{x}) - F(x) \leq -\frac{1}{2\tau}\|\mathcal{G}_\tau(x)\|^2. \tag{3.2}$$

To derive the linear convergence, we use (3.1) with $x = x_k, \tilde{x} = x_{k+1}$:

$$\begin{aligned} 0 &\leq \|\mathcal{G}_\tau(x_k)\|^2 \left(\frac{1}{\gamma} - \frac{1}{2\tau}\right) - F(x_{k+1}) + F(\bar{x}) \\ &= \frac{1}{2\tau}\|\mathcal{G}_\tau(x_k)\|^2 \left(\frac{2\tau}{\gamma} - 1\right) - F(x_{k+1}) + F(\bar{x}) \\ &\underset{(1)}{\leq} \left(\frac{2\tau}{\gamma} - 1\right)(F(x_k) - F(x_{k+1})) - F(x_{k+1}) + F(\bar{x}) \\ &= \left(\frac{2\tau}{\gamma} - 1\right)(F(x_k) - F(\bar{x}) + F(\bar{x}) - F(x_{k+1})) - F(x_{k+1}) + F(\bar{x}) \\ &= \frac{2\tau}{\gamma}(F(\bar{x}) - F(x_{k+1})) + \left(\frac{2\tau}{\gamma} - 1\right)(F(x_k) - F(\bar{x})). \end{aligned}$$

At (1) we used (3.2). Multiple both side by $\frac{\gamma}{2\tau}$ then we are done. ■

3.2 Conditions for linear convergence of the proximal problem

In this section, we will focus on the sufficient characterization of the proximal problem which allows proximal gradient method to achieve linear convergence rate. The following assumption characterizes a set of sufficient conditions of the proximal problem such that linear convergence rate of applying ISTA to dual proximal objective Ψ_λ can be achieved.

{ass:lin-cnvg-for-pp}
Assumption 3.4 (conditions for linear convergence of proximal problem) This assumption is about $(g, \omega, A, \Phi_\lambda, \Psi_\lambda, L_\Phi^\lambda, \gamma_\lambda)$. Here are the assumptions

(i) Assume (g, ω, A) satisfies Assumption 1.15. The primal objective Φ_λ is given by (1.1), and dual Ψ_λ by (1.2). This means if we let $f(v) := \frac{1}{2\lambda}\|\lambda A^\top v - y\|^2 - \frac{1}{2\lambda}\|y\|^2, g(v) := \omega^\star(v)$, then $\Psi_\lambda = f + g$.
(ii) Next, assume $\Psi_\lambda = f + g$ satisfies Assumption 3.1 by viewing $\Psi_\lambda = F$ and denote $\gamma_\lambda := \gamma$. Note that we can do this because f is obviously Lipschitz smooth.
(iii) Assume the primal objective Φ_λ is L_Φ^λ Lipschitz continuous.

The following propositions precisely show that the linear convergence is achievable when Assumption 3.4 holds.

{prop:inn-loop-lin-cnvg}
Proposition 3.5 (sufficient conditions of linear convergence of the inner loop) *Let the parameters $(g, \omega, A, \Phi_\lambda, \Psi_\lambda, L_\Phi^\lambda, \gamma_\lambda)$ of a proximal problem satisfy Assumption 3.4. Let $\tau \geq L_\Psi^\lambda$, and let $v_0 \in \operatorname{dom} \Psi_\lambda$ be an initial condition for an algorithm that satisfies Definition 3.2 with $F(v) = f(v) + g(v)$ and suppose that it generates iterates $(v_n)_{n\geq 0}$. Let $\bar{v}$ be a minimizer of Ψ_λ, then the followings are true for all $\epsilon \geq 0$:*

{prop:inn-looplin-cnvg-item1}
(i) *The sequence $\Psi_\lambda(v_n) - \Psi_\lambda(\bar{v})$ converges linearly to zero.*
{prop:inn-looplin-cnvg-item2}
(ii) *If we let $z_n := y - \lambda A^\top v_n$, and $\bar{z} := y - \lambda A^\top \bar{v}$, then $\Phi_\lambda(z_n) - \Phi_\lambda(\bar{z})$ converges linearly.*
{prop:inn-looplin-cnvg-item3}
(iii) *The duality gap has convergence $\mathbf{G}_\lambda(z_n, v_n) \leq \mathcal{O}\left(\left(1 - \frac{\gamma_\lambda}{2\tau}\right)^{n/2}\right)$.*
{prop:inn-looplin-cnvg-item4}
(iv) *If $\mathbf{G}_\lambda(z_n, v_n) \leq \epsilon$, then $z_n \approx_\epsilon \operatorname{prox}_{\lambda g}(y)$.*

Proof. To see (i), the sequence $\Psi_\lambda(v_n) - \Psi_\lambda(\bar{v})$ has linear convergence by Theorem 3.3 because we assumed that Ψ_λ satisfies Assumption 3.4. More precisely it has:

{ineq:inn-loop-lin-cnvg-pitem1}
$$\Psi_\lambda(v_n) - \Psi_\lambda(\bar{v}) \leq \left(1 - \frac{\gamma_\lambda}{2\tau}\right)^n \left(\Psi_\lambda(v_0) - \Psi_\lambda(\bar{v})\right). \tag{3.3}$$

We will now proceed to proving (ii). Because the iterates are given by $z_n = y - \lambda A^\top v_n$, and $\bar{z} = y - \lambda A^\top \bar{v}$, due assumed Lipschitz continuity of Φ_λ, we can use Theorem 1.18 which

gives:

{ineq:inn-loop-lin-cnvg-pitem2}

$$\begin{aligned}\Phi_\lambda(z_n) - \Phi_\lambda(\bar{z}) &\leq L_\Phi^\lambda \sqrt{2\lambda}\left(\Psi_\lambda(v_n) - \Psi_\lambda(\bar{v})\right)^{1/2} \\ &\underset{(1)}{\leq} L_\Phi^\lambda \sqrt{2\lambda}\left(1 - \frac{\gamma_\lambda}{2\tau}\right)^{n/2}\left(\Psi_\lambda(v_0) - \Psi_\lambda(\bar{v})\right)^{1/2}.\end{aligned} \tag{3.4}$$

At (1), we used (3.3).

To prove (iii), by definition of duality gap in (1.3) we have:

$$\begin{aligned}&\mathbf{G}_\lambda(z_n, v_n) \\ &\underset{(1)}{=} \Phi_\lambda(z_n) + \Psi_\lambda(v_n) + 0 \\ &= \Phi_\lambda(z_n) + \Psi_\lambda(v_n) - \Phi_\lambda(\bar{z}) - \Psi_\lambda(\bar{v}) \\ &= \Phi_\lambda(z_n) - \Phi_\lambda(\bar{z}) + \Psi_\lambda(v_n) - \Psi_\lambda(\bar{v}) \\ &\underset{(2)}{\leq} L_\Phi^\lambda \sqrt{2\lambda}\left(1 - \frac{\gamma_\lambda}{2\tau}\right)^{n/2}\left(\Psi_\lambda(v_0) - \Psi_\lambda(\bar{v})\right)^{1/2} + \left(1 - \frac{\gamma_\lambda}{2\tau}\right)^{n}\left(\Psi_\lambda(v_0) - \Psi_\lambda(\bar{v})\right) \\ &= \left(1 - \frac{\gamma_\lambda}{2\tau}\right)^{n/2}\left(L_\Phi^\lambda \sqrt{2\lambda}\left(\Psi_\lambda(v_0) - \Psi_\lambda(\bar{v})\right)^{1/2} + \left(1 - \frac{\gamma_\lambda}{2\tau}\right)^{n/2}\left(\Psi_\lambda(v_0) - \Psi_\lambda(\bar{v})\right)\right).\end{aligned}$$

At (1) we used strong duality because constraint qualification is in Assumption 1.15. At (2) we used (3.3), (3.4). To see (iv), apply Theorem 1.17. ■

Remark 3.6 In pratice, use the proximal operator of $\omega^\star$ choose a feasible v_0.

The next proposition will characterize a precise case where Assumption 3.4 is true, and it's a case widely available in applications.

{prop:1nrm-prox-problem}

Proposition 3.7 (1-norm problem) *Let (g, ω, A) satisfies Assumption 1.15. In addition, if $g := \|\cdot\|_1$, then the function $\Psi_\lambda, \Psi_\lambda$ satisfies Assumption 3.4.*

Proof. Take note that the dual has closed form $g^\star(z) = \delta(z|\{x : \|x\|_1 \leq 1\})$. The $g^\star$ is an indicator of a polytope constraint. The objective function $\Psi_\lambda(v) = \frac{1}{2\lambda}\|\lambda A^\top v - y\|^2 + \omega^\star(v) - \frac{1}{2\lambda}\|y\|^2$ so it's Lipschitz smooth, and it must have a set of minimizers S because its domain is compact. In addition, Ψ_λ fits the assumption of Necoara et al. [1, Theorem 8]. Therefore, Ψ_λ is quasi-strongly convex then by [1, Theorem 4], and [1, Theorem 7], it satisfies error bound condition as given in Assumption 3.1, hence Assumption 3.4 also.

Let $\mu_f = \|A\|^2/\kappa_f$, let $\kappa_f = \theta^{-2}(A, C)$ be the Hoffman Constant as presented by Necoara et al. in [1, Section 4] where C is the constraint matrix of the inequality set $\delta(z|\{x : \|x\|_1 \leq 1\})$, then the error bound constant can be satisfied with

$$\gamma_\lambda = \frac{\kappa_f}{1 + \lambda\mu_f + \sqrt{1 + \lambda\mu_f}}.$$

■

Remark 3.8 It is very difficult to obtain a lower estimate for γ in practice.

Definition 3.9 (piecewise linear-quadratic function [2, Definition 10.20]) *A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is piecewise linear-quadratic when* $\operatorname{dom} f$ *is the union of finitely many piecewise polyhedral set, such that on each polyhedral partition of f there exists $\alpha \in \mathbb{R}, a \in \mathbb{R}^n$, $C \in \mathbb{R}^{n\times n}$ as a symmetric matrix is where f has the representation $x \to \langle x, Cx\rangle + \langle a, x\rangle + c$.*

The following proposition characterizes another case where error bound conditions of problem holds.

{prop:plq-pp}

{prop:plq-prox-problem}

Proposition 3.10 (convex PLQ proximal problem)

{example:inn-lp-lin-cnvg}

Proposition 3.11 (scenario II, convex PLQ linear composite)

Example 3.12 (inner loop linear convergence rate examples)

3.3 Inner loop complexity

In this section, we will estimate the total number of iterations in the inner loop to achieve a given accuracy ϵ on primal objective Φ_λ of the proximal problem for some given $\lambda > 0$. We will specify the context for parameters ϵ, λ in relation to the outer problem of obtaining $x_k \approx_{\epsilon_k} T_{B_k+\rho_k}(y_k)$.

4 Total complexity of the algorithm

This section puts results regarding the total complexity of the proposed inexact proximal gradient algorithm.

References

[1] I. Necoara, Y. Nesterov, and F. Glineur, *Linear convergence of first order methods for non-strongly convex optimization*, Mathematical Programming, 175 (2019), pp. 69–107.

[2] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, vol. 317 of Grundlehren der mathematischen Wissenschaften, Springer, Berlin, Heidelberg, 1998.

[3] S. Villa, S. Salzo, L. Baldassarre, and A. Verri, *Accelerated and inexact forward-backward algorithms*, SIAM Journal on Optimization, 23 (2013), pp. 1607–1633.

[4] C. Zalinescu, *Convex analysis in general vector spaces*, World Scientific, River Edge, N.J. ; London, 2002.

[5] M. Zhang, M. Zhang, F. Zhang, A. Chaddad, and A. Evans, *Robust brain MR image compressive sensing via re-weighted total variation and sparse regression*, Magnetic Resonance Imaging, 85 (2022), pp. 271–286.