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

Author 1 Name, Author 2 Name *

November 6, 2025

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

Abstract

This is still a draft. [6].

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 Preliminaries

{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 [5, 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. [4, 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 $\tilde{y} \approx_\epsilon \operatorname{prox}_{\lambda^{-1}g^\star}(\lambda^{-1}y)$, then it has:

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

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. ■

Fact 1.9 (Fenchel Rockafellar duality)

1.2 Inexact proximal gradient inequality

{ass:for-inxt-pg-ineq}

Assumption 1.10 (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 1.10 has

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

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

{def:inxt-pg}

Definition 1.11 (inexact proximal gradient) *Let (f, g, L) satisfies Assumption 1.10. Let $\epsilon \ge 0, \rho > 0$. 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 1.12 We assumed that we can get exact evaluation of ∇f at any points $x \in \mathbb{R}^n$.

{lemma:other-repr-inxt-pg}

Lemma 1.13 (other representations of inexact proximal gradient) *Let (f, g, L) satisfies Assumption 1.10, $\epsilon \ge 0, \rho > 0$, 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 1.14 (inexact over-regularized proximal gradient inequality) *Let (F, f, g, L) satisfies Assumption 1.10, $\epsilon \ge 0, B \ge 0, \rho > 0$. Consider $\tilde x \approx_\epsilon T_{B+\rho}(x)$. Denote $F = f + g$. If in addition, $\tilde x, B$ satisfies the line search condition $D_f(\tilde x, x) \le B/2\|x - \tilde x\|^2$, then it has $\forall z \in \mathbb{R}^n$:*

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

Proof. By Definition 1.11 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 $z \in \mathbb{R}^n$:

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

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

Remark 1.15 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 $\approx_\epsilon$ 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 $\tilde{x} \approx_\epsilon \mathrm{prox}_{\lambda g}(z)$. Eventually we want to evaluate $T_\rho(x)$ of some $F = f + g$ inexactly using Lemma 1.13. To do that one would need to evaluate $\mathrm{prox}_{\rho^{-1} g}$ 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 1.16 (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 $\mathrm{prox}_{\lambda\omega}$ and, its conjugate $\omega^\star$ is known.
(iii) $g := \omega(Ax)$ such that $\mathrm{rng}\, A \cap \mathrm{ri}\,\mathrm{dom}\, g \neq \emptyset$.

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

{eqn:primal-pp}

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

Observe that $\operatorname{rng} A \cap \operatorname{ri} \operatorname{dom} g \neq \emptyset$ in Assumption 1.16 shows g is also closed convex and proper. The function Φ_λ is coersive due to its quadratic term and hence it must admit a minimizer [3, Theorem 1.9]. Recall the following famous theoretical result in the convex programming literature that we had adapted into our context.

{fact:fn-rck-duality} **Fact 1.17 (French Rockafellar Duality [1, Proposition 15.22])** *Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be closed convex and proper, $g : \mathbb{R}^m \to \overline{\mathbb{R}}$, $A \in \mathbb{R}^{m\times n}$. If $\mathbf{0} \in \operatorname{int}(\operatorname{dom} g - A \operatorname{dom} f)$, then*

$$\inf_{u\in\mathbb{R}^n} \{f(u) + g(Au)\} + \min_{v\in\mathbb{R}^m} \left\{f^\star \circ (-A^\top) + g^\star(v)\right\} = 0.$$

Remark 1.18 The theorem is not exactly the same as what is claimed in the original text, since we are in a finite dimensional setting. To see the original theorem cited in the finite dimension, we need the space $\mathcal{H} = \mathbb{R}^n$ and then use [1, Proposition 6.12].

In our context, we are interested in the dual of the proximal problem Φ_λ which makes $f = u \mapsto \frac{1}{2\lambda}\|u - y\|^2$, and $g = \omega$, and it has $f^\star(v) = \frac{1}{2\lambda}\|\lambda v + y\|^2 - \frac{1}{2\lambda}\|y\|^2$ (see Apendix A.1). Consequently, $[f^\star \circ (-A^\top)](v) = \frac{1}{2\lambda}\| - \lambda A^\top v + y\|^2 - \frac{1}{2\lambda}\|y\|^2$. And therefore, Φ_λ admits Fenchel Rockafellar dual (or simply the dual) objective in $\mathbb{R}^m$:

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

We define the duality gap

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

Note that in this case the smooth part is quadratic and, $\operatorname{dom} f = \mathbb{R}^n$, hence it translates to $\mathbf{0} \in \operatorname{int}(\operatorname{dom} g - A \operatorname{dom} f) = \operatorname{int}(\operatorname{dom} g - \operatorname{rng} A)$. This will hold because of $\operatorname{rng} A \cap \operatorname{ri} \operatorname{dom} g \neq \emptyset$ in Assumption 1.16. Therefore strong duality holds and it exists $(\hat u, \hat v)$ such that we have the following:

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

The following theorem quantifies a sufficient conditions for $\tilde x \approx_\epsilon \operatorname{prox}_{\lambda g}(x)$. The theorem below is from [4, Proposition 2.2].

{thm:primal-dual-trans} **Theorem 1.19 (primal translate to dual [4, Proposition 2.2])** *Let (g, ω, A) satisfies assumption 1.16, $\epsilon \geq 0$, then*

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

This theorem that follows is from Villa et al. [4, Proposition 2.3].

{thm:dlty-gap-inxt-pp} **Theorem 1.20 (duality gap of inexact proximal problem [4, Proposition 2.3])** *Let (g, ω, A) satisfies Assumption 1.16, for all $\epsilon \ge 0$, $v \in \mathbb{R}^n$ consider the following conditions:*

(i) $\mathbf{G}_\lambda(y - \lambda A^\top v, v) \le \epsilon$.
(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. [4, Proposition 2.3], and to show $(b) \iff (c)$ use Theorem 1.8. ■

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. [4, Theorem 5.1].

{fact:minimizing-dual-pp} **Fact 1.21 (minimizing dual of the proximal problem [4, Theorem 5.1])** *Let $\bar{v}$ be a solution of Ψ_λ. Suppose that $(v_n)_{n\ge 0}$ is a minimizing sequence for Ψ_λ. Let $z_n = y - \lambda A^\top v_n$, and $\bar{z} = y - \lambda A^\top \bar{v}$. Then the following are true:*

(i) *The sequence $z_n \to \bar{z}$ as well.*
(ii) *There exists a constant K that depends on the sequence z_n, and Φ_λ such that for all $k \ge 0$ the inequality:*

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

We remark that the above fact translates any algorithm that optimizes the dual problem Ψ_λ into optimizing the duality gap $\mathbf{G}(z_n, v_n)$ because it shows the primal is bounded by the dual through a Lipschitz-like inequality.

With the proof back in Villa et al. [4, Theorem 5.1], pay close attention to the constant L_Φ exists by the virtue of Ψ being convex, and the sequence $v_n \to \bar{v}$ which allows the sequence $z_n \to \bar{z}_n$ to be bounded inside the domain of Φ_λ. By convexity, Φ_λ is locally Lipschitz on $\operatorname{ri}\operatorname{dom}\Phi_\lambda$, by convergence of z_n the sequence is in a compact set in $\operatorname{ri}\operatorname{dom}\Phi_\lambda$ hence a Lipschitz bound with constant K exists.

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 1.20 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 1.11, 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 1.10, 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:

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

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 $k \ge 1$ 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 $k \ge 1$ 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 1.10, 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 1.14 has:

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

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:

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

■

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.

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

Therefore, it points to the following inequality:

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

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

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

■

{prop:apg-cnvg-gm-generic} **Proposition 2.5 (generic convergence rate of the gradient mapping)**

Cette proposition ira démontrer que le séquence de "Inexact Gradient Mapping" converge.

Proof.

■

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 $(\alpha_k)_{k\ge 0}$.

2.1 Convergence results of the outer loop

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

{ass:opt-mmntm-seq} **Assumption 2.6 (the optimal momentum sequence)**
Keeping everything assumed in Assumption 2.3 about $(\alpha_k, B_k, \rho_k, \epsilon_k)_{k\ge 0}$, (F, f, g, L), $(y_k, x_k, v_k)_{k\ge 0}$, $(\beta_k)_{k\ge 0}$, $\mathcal{E}_0$ and p. We assume in addition that the sequence $(\alpha_k)_{k\ge 0}$ satisfies for all $k \ge 0$ the equality: $(1 - \alpha_k) = \alpha_k^2 L_k \alpha_{k-1}^{-2} L_{k-1}^{-1}$ and, $p > 1$.

{lemma:opt-mmntm-seq} **Lemma 2.7 (the optimal momentum sequence is indeed valid and optimal)**
Let $(\alpha_k)_{k\ge 0}, (\beta_k)_{k\ge 0}$ be given by Assumption 2.6. If we choose $\alpha_0 \in (0, 1]$ then for all $k \ge 1$

it has:

{eqn:lemma:opt-mmntm-seq-result1}
$$\alpha_k = \frac{L_{k-1}}{2L_k}\left(-\alpha_{k-1}^2 + \left(\alpha_{k-1}^4 + 4\alpha_{k-1}\frac{L_k}{L_{k-1}}\right)^{1/2}\right) \in (0,1) \tag{2.2}$$

For the sequence $(\beta_k)_{k\geq 0}$ *it has* $\forall k \geq 1$

{ineq:lemma:opt-mmntm-seq-result2}
$$\left(1 + \alpha_0\sqrt{L_0}\sum_{i=1}^{k}\sqrt{L_i^{-1}}\right)^{-2} \leq \beta_k \leq \left(1 + \frac{\alpha_0\sqrt{L_0}}{2}\sum_{i=1}^{k}\sqrt{L_i^{-1}}\right)^{-2}. \tag{2.3}$$

Proof. Firstly, we will show (2.2). We will prove using induction. Fix any $k \geq 1$. Assume inductively that $\alpha_{k-1} \in (0,1]$. We can solve for α_k using the recursive equality $(1-\alpha_k) = \alpha_k^2 L_k \alpha_{k-1}^{-2} L_{k-1}^{-1}$ from Assumption 2.6. Writing α_{k-1} as α, and L_k/L_{k-1} as q. We then solve for α_k, the quadratic equation always admits one root that is strictly positive which is given as:

$$\begin{aligned}
\alpha_k &= \frac{1}{2}\left(-\frac{\alpha^2}{q} + \sqrt{\frac{\alpha^4}{q^2} + \frac{4\alpha^2}{q}}\right) \\
&= \frac{\alpha^2}{2q}\left(-1 + \sqrt{1 + \frac{4q}{\alpha^2}}\right) \\
&\underset{(1)}{<} \frac{\alpha^2}{2q}\left(-1 + 1 + \frac{2q}{\alpha^2}\right) \\
&= 1
\end{aligned}$$

At (1) we completed a square in the radical, and we used the assumption $\alpha_k > 0$ and, $L_k > 0, L_{k-1} > 0$ because we had $B_k > 0$, therefore the following chain of inequality holds:

$$\begin{aligned}
1 + \frac{4q}{\alpha^2} &= 1 + \frac{4q}{\alpha^2} + \frac{4q^2}{\alpha^4} - \frac{4\alpha^2}{\alpha^2} \\
&= \left(1 + \frac{2q}{\alpha^2}\right)^2 - \frac{4q^2}{\alpha^4} \\
&< \left(1 + \frac{2q}{\alpha^2}\right)^2.
\end{aligned}$$

To see that $\alpha_k > 0$, recall the same fact that $L_k > 0$, and the inductive hypothesis $\alpha_{k-1} \in (0,1]$ then $4q/\alpha^2 > 0$ so obviously $\alpha_k = \frac{\alpha^2}{2q}\left(-1 + \sqrt{1 + 4q/\alpha^2}\right) > 0$ because the quantity inside the radical is strictly larger than 1. Therefore, inductively it holds that $\alpha_k \in (0,1)$ too.

We will now show (2.3). 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}

$$\begin{aligned}
&(1-\alpha_k) = \beta_k/\beta_{k-1} \\
\iff\ & \alpha_k = 1 - \beta_k/\beta_{k-1} \\
\implies\ & \alpha_k^2 = (1-\beta_k/\beta_{k-1})^2.
\end{aligned} \tag{2.4}$$

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

{eqn:opt-mmntm-seq-pitem2}

$$\begin{aligned}
\alpha_k^2 &= (1-\alpha_k)\alpha_{k-1}^2 L_{k-1} L_k^{-1} \\
&= (1-\alpha_k)\left(\frac{\alpha_{k-1}^2 L_{k-1}}{\alpha_0^2 L_0}\right)\frac{\alpha_0^2 L_0}{L_k} \\
&= \left(\beta_k \beta_{k-1}^{-1}\right)\left(\beta_{k-1}\right) L_0 \alpha_0^2 L_k^{-1} \\
&= \beta_k L_0 \alpha_0^2 L_k^{-1}.
\end{aligned} \tag{2.5}$$

Combining (2.4), (2.5) 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 for $i = 1, 2, \ldots, k$, use the fact that $\beta_0 = 1$ will yield the desired results after some algebraic manipulations. ■

Remark 2.8

{prop:opt-cnvg-outr-loop} **Proposition 2.9** **($\mathcal{O}(1/k^2)$ 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 2.6. Assume in addition that*

(i) *there exists $\bar{x} \in \mathbb{R}^n$ that is a minimizer of $F = f + g$;*
(ii) *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, 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 Lemma 2.7 because of the same assumption on $(\alpha_k)_{k\geq 0}$, then using the fact that L_k is bounded:

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

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

Remark 2.10 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 some parameters used in the inexact algorithm.

Pay attention that α_k had been determined in the above theorem (as seemed in Lemma 2.7), 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 $(\rho_k)_{k\geq 0}$. Since we only need the entire sequence $L_k = \rho_k + B_k$ 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 counters ϵ_k in the proximal gradient inequality, 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 $(\epsilon_k)_{k\ge 0}$ 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 2.9 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 2.11 (error schedule lower bound)**
Let, $(\alpha_k, B_k, \rho_k, \epsilon_k)_{k\ge 0}$, $(\beta_k)_{k\ge 0}, \mathcal{E}_0, p$ be given by Assumption 2.6, $L_k := \rho_k + B_k$. Let $(\epsilon_k)_{k\ge 0}$ 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 $k \ge 0$ such that it has $L_{\min} \le \min_{1\le i\le k} L_i$ and, we assume $\mathcal{E} > 0$, then ϵ_k admits the non-trivial lower bound:

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

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

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

At (1), we used Lemma 2.7. At (2), we used that $L_{\min} \le L_i$ 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 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's assume that the outer loop iteration k is fixed throughout this entire section. Now 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.

{def:pg-and-gm} **Definition 3.1 (the exact proximal gradient and gradient mapping)**
Let $F = f + g$ where $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function and, $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is closed convex and proper. We define the exact proximal gradient mapping and, gradient mapping:

(i) *The proximal gradient operator $T_\tau(x) := \operatorname{prox}_{\tau^{-1}g}(x - \tau^{-1}\nabla f(x))$,*
(ii) *and the gradient mapping operator $\mathcal{G}_\tau(x) := \tau(x - T_\tau(x))$.*

{def:err-bnd} **Definition 3.2 (error bound)**
Let $F = f + g$ where $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function and, $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is closed convex and proper. Assume the set of minimizers $S = \operatorname*{argmin}_x F \neq \emptyset$. Let $\gamma > 0$, and let proximal gradient mapping T_τ, gradient mapping $\mathcal{G}_\tau$ of $F = f + g$ as given by Definition 3.1. Then their sum $F = f + g$ satisfies the error bound condition if there exists $\gamma > 0$ such that

{ineq:pg-eb}
$$\|\mathcal{G}_\tau(x)\| \geq \gamma \operatorname{dist}(x|S). \tag{3.1}$$

{ass:pg-eb} **Assumption 3.3 (gradient mapping error bound)**
The following assumption is about (F, f, g, L, S, γ). Assume that

(i) (f, g, L) satisfies Assumption 1.10,
(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 $F = f + g$ and, it satisfis error bound (Definition 3.2).

{def:ista} **Definition 3.4 (proximal gradient method)**
Suppose that (f, g, L) satisfies Assumption 1.10. 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 3.5 (linear convergence under gradient mapping error bound)** *Assume that (F, f, g, L, S, γ) is given by Assumption 3.3. Under this assumption, the iterates $(x_k)_{k\geq 0}$ given by Definition 3.4 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 1.14, 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 3.6 $\exists\gamma > 0$ 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}

$$0 \leq F(\tilde{x}) - F(\bar{x}) \leq \|\mathcal{G}_\tau(x)\|^2\left(\frac{1}{\gamma} - \frac{1}{2\tau}\right). \tag{3.2}$$

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

{ineq:lin-cnvg-ista-eb-pitem2}

$$F(\tilde{x}) - F(x) \le -\frac{1}{2\tau}\|\mathcal{G}_\tau(x)\|^2. \tag{3.3}$$

To derive the linear convergence, we use (3.2) with $x = x_k, \tilde{x} = x_{k+1}$:

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

At (1) we used (3.3). Multiple both side by $\frac{\gamma}{2\tau}$ then we are done. ■

{sec:conds-lin-cnvg-pp}

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 3.6 (conditions for linear convergence of proximal problem) This assumption is about $(g, \omega, A, y, \Phi_\lambda, \Psi_\lambda, L_\Phi^\lambda, \gamma_\lambda)$. Here are the assumptions

(i) $y \in \mathbb{R}^n$ is the vector of which the proximal problem is anchored at, fixed it to be an arbitrary vector.

(ii) Assume (g, ω, A) satisfies Assumption 1.16. The primal objective Φ_λ is given by (1.1), and dual Ψ_λ by (1.2). This means if we let $h(v) := \frac{1}{2\lambda}\|\lambda A^\top v - y\|^2 - \frac{1}{2\lambda}\|y\|^2$, then $\Psi_\lambda(v) = h(v) + \omega^\star(v)$.

(iii) Next, assume $\Psi_\lambda = h + \omega^\star$ satisfies error bound condition (Assumption 3.3) with $\Psi_\lambda = F$ and, $\gamma = \gamma_\lambda$ and $f = h$. Note that we can do this because h is quadratic hence obviously Lipschitz continuous and Lipschitz smooth.

(iv) We assume in addition, the primal objective Φ_λ of proximal problem is L_Φ^λ Lipschitz continuous on its domain $\operatorname{dom}\Phi_\lambda = \operatorname{dom}(g \circ A)$.

The following definition specifies the algorithm that can achieve linear convergence rate with the assumptions above.

{def:ista-inner-lp} **Definition 3.7 (the ISTA inner loop algorithm)**
Let $\lambda > 0, \epsilon > 0$, and $(g, \omega, A, y, \Phi_\lambda, \Psi_\lambda, L_\Phi^\lambda, \gamma_\lambda)$ satisfies Assumption 3.6.

(i) *Let $v_0 \in \operatorname{dom} \omega^\star$ be a feasible initial guess of Ψ_λ, and let $\tau \geq \lambda \|AA^\top\|$ be the inverse step size.*
(ii) *Define $z_0 = y - \lambda A^\top v_0$ and smooth part of Ψ_λ as $h := v \mapsto \frac{1}{2\lambda}\|\lambda A^\top v - y\|^2$.*

The algorithm that solves the proximal problem generates the primal dual sequences (z_j, v_j) such that for all $j = 0, 1, 2, \cdots$, they satisfy:

$$
\begin{aligned}
v_{j+1} &= \operatorname{prox}_{\tau^{-1}\omega^\star}\left(v_j - \tau^{-1}\nabla h(v_j)\right), \\
z_{j+1} &= y - \lambda A^\top v_{j+1}.
\end{aligned}
$$

Terminates if $\mathbf{G}_\lambda(z_j, v_j) \leq \epsilon$ where $\mathbf{G}_\lambda$ is given by (1.3).

Remark 3.8 The value of $\mathbf{G}_\lambda(z_j, v_j)$ is easy to compute, it only requires access to matrix $A, A^\top$, and the function ω. In case when the proximal operator for $\omega^\star$ is nontrivial, we can use the Moreau identity and the proximal operator of ω instead. The gradient $\nabla f(v)$ is easy to compute, and it is: $AA^\top v - Ay$.

The following propositions precisely show that the linear convergence is achievable when Assumption 3.6 holds.

{prop:inn-loop-lin-cnvg} **Proposition 3.9 (sufficient conditions of linear convergence of the inner loop)**
Let the parameters $(g, \omega, A, y, \Phi_\lambda, \Psi_\lambda, L_\Phi^\lambda, \gamma_\lambda)$ of a proximal problem satisfy Assumption 3.6. Let τ, v_0, ϵ be given by an algorithm that satisfies Definition 3.7 and, it generates iterates $(z_j, v_j)_{j\geq 0}$. Let $\bar{v}$ be a minimizer of Ψ_λ, then the followings are true:

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

Proof. Note that Ψ_λ satisfies error bound condition in Assumption 3.6, and Definition 3.7 has $\tau \geq \lambda\|A^T A\|$ which satisfies Definition 3.4 therefore the linear convergence results of Theorem 3.5 applies to $\Psi_\lambda(v_j) - \Psi_\lambda(\bar{v})$. Therefore, we claim:

{ineq:inn-loop-lin-cnvg-pitem1}
$$
\Psi_\lambda(v_j) - \Psi_\lambda(\bar{v}) \leq \left(1 - \frac{\gamma_\lambda}{2\tau}\right)^j \left(\Psi_\lambda(v_0) - \Psi_\lambda(\bar{v})\right). \tag{3.4}
$$

We will now proceed to proving (ii). Recall that the iterates $z_j = y - \lambda A^\top v_j$, and $\bar{z} = y - \lambda A^\top \bar{v}$, and we have Lipschitz continuity of Φ_λ, hence Theorem 1.21 applies, in addition since Φ_λ were assumed to be globally Lipschitz on $\operatorname{dom}\Phi_\lambda$ in Assumption 3.6, this means that we can replace K regardless of $\bar{v}$, and convergent primal sequence $z_j \to \bar{z}$ with L_Φ^λ which gives:

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

$$\begin{aligned}\Phi_\lambda(z_j) - \Phi_\lambda(\bar{z}) &\le L_\Phi^\lambda \sqrt{2\lambda}\left(\Psi_\lambda(v_j) - \Psi_\lambda(\bar{v})\right)^{1/2} \\ &\underset{(1)}{\le} L_\Phi^\lambda \sqrt{2\lambda}\left(1 - \frac{\gamma_\lambda}{2\tau}\right)^{j/2}\left(\Psi_\lambda(v_0) - \Psi_\lambda(\bar{v})\right)^{1/2}.\end{aligned} \tag{3.5}$$

At (1), we used (3.4).

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

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

At (1) we used strong duality because constraint qualification is in Assumption 1.16. At (2) we used (3.4), (3.5). To see (iv), apply Theorem 1.20. ■

Remark 3.10 In practice, use the proximal operator of $\omega^\star$ to choose a feasible v_0, equivalently we can translate any primal feasible solution into a dual feasible solution using Theorem 1.19.

As a prelude, to have a global convergence rate it requires an upper bound of the initial optimally gap of Ψ_λ of for all iterations of the outer loop. Suppose for each outer iteration k, the inner loop is initialized to start on $v_0^{(k)}$, then there must be an upper bound for all $\Psi_\lambda\left(v_0^{(k)}\right) - \Psi_\lambda(\bar{v})$.

3.3 Concrete examples where we can have linear convergence

Some results in the literature now follow. The goal of this section is to show that specific type of nonsmooth function ω that common appear in applications satisfies the error bound condition for the proximal problem, and hence the results from the previous section are applicable.

Definition 3.11 (quasi-strongly convex function [2, Definition 1]) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable. Let $X \subseteq \mathbb{R}^n$ and suppose that the set of minimizers $X^+ = \operatorname*{argmin}_{x\in X} f(x) \neq \emptyset$ exists and, denote f^+ to be the minimum of f on X. It is quasi strongly-convex on the set $X \subseteq \mathbb{R}^n$ if $\exists \kappa > 0$ such that $\forall x \in X$, let $\bar{x} = \Pi_{X^+}x$, it satisfies*

$$f^+ - f(x) - \langle \nabla f(x), \bar{x} - x\rangle + \frac{\kappa}{2}\|x - \bar{x}\|^2 \quad (\forall x \in X).$$

{def:q-growth}

Definition 3.12 (quadratic growth) *Let $F = f + g$ where $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and, $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be closed convex and, proper. Assume in addition that the set of minimizer $S := \operatorname*{argmin}_{x\in\mathbb{R}^n} F \neq \emptyset$, denote F^+ to be the minimum, then it satisfies the quadratic growth conditions if there exists $\kappa > 0$ such that*

$$(\forall x \in \mathbb{R}^n)\ F(x) \geq F^+ + \frac{\kappa}{2}\operatorname{dist}^2(x|S).$$

The following theorem shows that under the standard framework of smooth nonsmooth additive compsitive objective function, the quadratic growth condition implies error bound.

{thm:qfg-means-eb} **Theorem 3.13 (quadratic growth implies error bound)**
Let $f : \mathbb{R}^n \to$

Proof. Denote $x^+ = T_\tau(x)$, $\bar{x} = \Pi_S x$, and $\bar{x}^+ = \Pi_S x^+$. We use the exact proximal gradient inequality by applying Theorem 1.14 with $\epsilon = 0, \rho = 0$ and, $z = \bar{x}^+$ gives:

$$\begin{aligned}
& 0 \geq F(x^+) - F(\bar{x}^+) - \tau\langle \bar{x}^+ - x^+, x^+ - x\rangle - \frac{\tau}{2}\|x - x^+\|^2 \\
& \quad \underset{(1)}{\geq} \frac{\kappa}{2}\|x^+ - \bar{x}^+\|^2 - \|\tau(\bar{x}^+ - x^+)\|\|x^+ - x\| - \frac{1}{2\tau}\|\tau(x - x^+)\|^2 \\
& \quad = \frac{\kappa}{2}\left\|x^+ - \bar{x}^+\right\|^2 - \frac{\tau}{2}\left(2\left\|\bar{x}^+ - x^+\right\|\left\|x^+ - x\right\| + \|x - x^+\|^2\right) \\
& \quad \underset{(2)}{=} \frac{\kappa + \tau}{2}\left\|x^+ - \bar{x}^+\right\|^2 - \frac{\tau}{2}(\|\bar{x}^+ - x^+\| + \|x^+ - x\|)^2 \\
& \quad \underset{(3)}{\geq} \frac{\kappa + \tau}{2}\left(\|x - \bar{x}\| - \|x - x^+\|\right)^2 - \frac{\tau}{2}(\|x - \bar{x}\|)^2 \\
\iff & 0 \geq \sqrt{\kappa + \tau}\left(\|x - \bar{x}\| - \|x - x^+\|\right) - \sqrt{\tau}(\|x - \bar{x}\|) \\
& \quad = (\sqrt{\kappa + \tau} - \sqrt{\tau})\|x - \bar{x}\| - \sqrt{\kappa + \tau}\|x - x^+\|.
\end{aligned}$$

At (1), we used quadratic growth assumption and substitute the inequality of Definition 3.12, and the Cauchy inequality. At (2), consider substituting $a = \bar{x}^+ - x^+, b = x^+ - x$, so the second terms with the parenthesis has inside $2\|a\|\|b\| + \|b\|^2 = (\|a\| + \|b\|)^2 - \|a\|^2$. At (3), we used the properties of projecting onto the set set of minimizer S, which gives:

$$
\begin{aligned}
&\|x - \bar{x}\| \le \|x - \bar{x}^+\| \le \|x - x^+\| + \|x^+ - \bar{x}^+\| \\
\implies &\|x - \bar{x}\| - \|x - x^+\| \le \|x^+ - \bar{x}^+\|
\end{aligned}
$$

Finally, re-arranging the results it should yield the following inequality:

$$
\begin{aligned}
\|\mathcal{G}_\tau(x)\| &= \|\tau(x - x^+)\| \\
&\ge \tau \left(\frac{\sqrt{\kappa + \tau} - \sqrt{\tau}}{\sqrt{\kappa + \tau}}\right) \|x - \bar{x}\| \\
&= \frac{\tau\kappa}{\kappa + \tau + \sqrt{\tau(\kappa + \tau)}} \|x - \bar{x}\|.
\end{aligned}
$$

The above is error bound conditions as defined in Definition 3.2, with $\gamma = \frac{\tau\kappa}{\kappa+\tau+\sqrt{\tau(\kappa+\tau)}}$. ■

Remark 3.14 This theorem is not exactly a new result. We adapted claim and the proof here so the notations stays consistent and to make it self contained.

Fact 3.15 (Hoffman error bound)

Fact 3.16 (a type of quasi strongly convex function)

The next proposition will characterize a precise case where Assumption 3.6 is true, and it's a case widely available in applications.

{prop:1nrm-prox-problem}

Proposition 3.17 (1-norm problem) *Let (g, ω, A) satisfies Assumption 1.16. In addition, if $g := \|\cdot\|_1$, then the function $\Psi_\lambda, \Phi_\lambda$ satisfies Assumption 3.6.*

Proof. To verify, it requires that

(i) The function Ψ_λ satisfies the error bound condition as specified in Assumption 3.3;
(ii) Ψ_λ is also a Lipschitz continuous function.

Take note that the dual has closed form $g^\star(z) = \delta(z|\{x : \|x\|_1 \le 1\})$. 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 continuous on its domain, 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. [2, Theorem 8]. Therefore, Ψ_λ is quasi-strongly convex then by [2, Theorem 4], and [2, Theorem 7], it satisfies error bound condition as given in Assumption 3.3, hence Assumption 3.6 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 [2, Section 4] where C is the constraint matrix of the inequality set $\delta(z|\{x : \|x\|_1 \le 1\})$, then the error bound constant can be satisfied with

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

■

Remark 3.18 It is very difficult to obtain a lower estimate for γ in practice.

Definition 3.19 (piecewise linear-quadratic function [3, 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} **Proposition 3.20 (convex PLQ proximal problem)**

{prop:plq-prox-problem} **Proposition 3.21 (scenario II, convex PLQ linear composite)**

{example:inn-lp-lin-cnvg} **Example 3.22 (inner loop linear convergence rate examples)**

4 Total complexity of the algorithm

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

4.1 Inner loop complexity

We remind that the parameters in the inner loop will change depending on the outer loop's iteration. To discuss the convergence we unfortunately have to re-introduce the proximal problem in the context of the accelerated proximal gradient method of the outer loop.

Let (F, f, g, L) and sequence $(\alpha_k, B_k, \rho_k, \epsilon_k)_{k\ge0}$ satisfies Definition 2.1. Fix any $k \ge 0$ to be the iteration counter of the outer loop. Let (g, ω, A) satisfies Assumption 1.16. The inner loop is an algorithm that solves the inexact proximal gradient problem $x \approx_{\epsilon_k} T_{B_k+\rho_k}(y_k)$ which is equivalent to evaluating:

$$x_k \approx_{\epsilon_k} \operatorname{prox}_{L_k^{-1} g}(y_k - L_k^{-1}\nabla f(y_k)).$$

Where, $L_k := B_k + \rho_k$. Let $\lambda^{(k)} := L_k^{-1}, \tilde{y}_k := y_k - L_k^{-1}\nabla f(y_k)$, the proximal problem boils down to optimizing the following function $\Phi_\lambda^{(k)}$ as defined by:

{eqn:primal-pp-k}

$$\Phi_\lambda^{(k)}(u) := \frac{1}{2\lambda^{(k)}}\|u - \tilde{y}_k\|^2 + \omega(Au). \tag{4.1}$$

And its dual which is

{eqn:dual-pp-k}

$$\Psi_\lambda^{(k)}(v) := \frac{1}{2\lambda}\left\|\lambda^{(k)}A^\top v - \tilde{y}_k\right\|^2 + \omega^\star(v) - \frac{1}{2\lambda^{(k)}}\|\tilde{y}_k\|^2. \tag{4.2}$$

The primal dual gap is $\mathbf{G}_\lambda^{(k)}(u,v) = \Phi_\lambda^{(k)}(u) + \Psi_\lambda^{(k)}(v)$. The above are identical to proximal problem defined in Section 1.3 except for the introduction of iteration counter k from the outer loop, and we had specified $\tilde{y}_k$ in relation to the outer loop. Finally, the inner loop algorithm is responsible for optimizing the optimality gap, it satisfies that $\mathbf{G}_\lambda^{(k)}(u,v) \le \epsilon_k$ and all results from Section 3.2 applies.

The following assumption specifies conditions where the complexity of the inner using ISTA can be globally bounded for all iteration of the accelerated proximal gradient algorithm of the outer loop. This is crucial for analyzing the global convergence rate of the algorithm.

{ass:bounded-inn-cmplx}

Assumption 4.1 (globally bounded inner loop complexity) For any integer $k \ge 0$, this assumption is about parameters of the proximal problem $\left(g, \omega, A, \tilde{y}_k, \Phi_\lambda^{(k)}, \Psi_\lambda^{(k)}, L_\Phi^{\lambda,(k)}, \gamma_\lambda^{(k)}\right)$ introduced at the beginning of this section, iterates $\left(z_j^{(k)}, v_j^{(k)}\right)_{j\ge 0}$, the primal dual solutions $\left(\bar{z}^{(k)}, \bar{v}^{(k)}\right)$ and additional constants $\lambda^{\min}, \lambda^{\max}, \gamma^{\min}, C_0$. The assumptions now follow.

(i) As specified in this section, the parameter $\lambda^{(k)} = (B_k + \rho_k)^{-1}$ are from the outer loop. We denote $L_k = B_k + \rho_k$ for short. We assume in addition, there exists $\lambda^{\min}, \lambda^{\max}$ such that:

$$-\infty < \lambda^{\min} \le \inf_{k \in \mathbb{N}\cup\{0\}} \lambda^{(k)} \le \sup_{k \in \mathbb{N}\cup\{0\}} \lambda^{(k)} \le \lambda^{\max} < \infty.$$

(ii) There exists a constant C_0 such that all the initial guesses given by an algorithm in the outer loop $v_0^{(k)}$ and minimizers $\bar{v}^{(k)}$ satisfies $C_0 = \sup_{k\in\mathbb{N}\cup\{0\}} \left\|v_0^{(k)} - \bar{v}^{(k)}\right\| < \infty$.

(iii) There exists the smallest error bound constant $\gamma^{\min}$ such that it lower bound all the constants for the proximal problem, i.e: $\exists \gamma_{\min} :\ 0 < \gamma^{\min} \le \inf_{k\in\mathbb{N}\cup\{0\}} \gamma_\lambda^{(k)}$.

(iv) $\tilde{y}_k, \Phi_\lambda^{(k)}, \Psi_\lambda^{(k)}, L_\Phi^{\lambda,(k)}, \gamma_\lambda^{(k)}$ all satisfies error bound conditions (Assumption 3.6) with $y = \tilde{y}_k, \Phi_\lambda = \Phi_\lambda^{(k)}, \Psi_\lambda = \Psi_\lambda^{(k)}, L_\Phi^\lambda = L_\Phi^{\lambda,(k)}$ and $\gamma_\lambda = \gamma_\lambda^{(k)}$.

(v) The iterates of the inner loop $(z_j^{(k)}, v_j^{(k)})_{k\geq 0}$ are produced by an algorithm that satisfies Definition 3.7.

The proposition that follows characterize the bare minimum requirements so that the inner loop's complexity depends only on required accuracies ϵ, and parameter λ for all initial guess of the outer loop.

{prop:inner-lp-cmplx} **Proposition 4.2 (inner loop complexity can be bounded globally)** *Let* $\left(g, \omega, A, \tilde{y}_k, \Phi_\lambda^{(k)}, \Psi_\lambda^{(k)}, L_\Phi^{\lambda,(k)}, \gamma_\lambda^{(k)}\right)$, $\left(z_j^{(k)}, v_j^{(k)}\right)_{j\geq 0}$, $\left(\bar{z}^{(k)}, \bar{v}^{(k)}\right)$ *and additional parameters* $\lambda^{\min}, \lambda^{\max}, \gamma^{\min}, C_0$ *such that they satisfy Assumption 4.1.*

{prop:inner-lp-cmplx-result1}
{prop:inner-lp-cmplx-result2}

(i) *If* $\lambda_1 > \lambda_2$ *then* $\Psi_{\lambda_1}(v) \geq \Psi_{\lambda_2}(v)$ *for all* $v \in \operatorname{dom}\omega^\star$.

(ii) *As a consequence of the former, the sequence* $L_\Phi^{\lambda,(k)}$ *as specified back in Fact 1.21 is bounded above so, there exists* $L_\Phi^{\max}$ *such that* $L_\Phi^{\max} = \sup_{k\in\mathbb{N}\cup\{0\}} L_\Phi^{\lambda,(k)} < \infty$.

{prop:inner-lp-cmplx-result3}

(iii) *There exists a constant* C_1 *that bounds the optimality gap of all initial guesses:*

$$\infty > C_1 \geq \sup_{k\in\mathbb{N}\cup\{0\}} \left\{\Psi_\lambda^{(k)}(v_0^{(k)}) - \Psi_\lambda^{(k)}(\bar{v}^{(k)})\right\}.$$

{prop:inner-lp-cmplx-result4}

(iv) *There exists a constant* $\infty > C_2 \geq \sqrt{C_1}\sup_{k\in\mathbb{N}\cup\{0\}}\left\{L_\Phi^{\lambda,(k)}\sqrt{2\lambda^{(k)}} + \sqrt{C_1}\right\}$.

{prop:inner-lp-cmplx-result5}

(v) *Denote* N_ϵ *to be the total number of iteration of inner loop needed to achieve* $\mathbf{G}_\lambda^{(k)}(z_j, v_j) \leq \epsilon$, *then it's sufficient to have*

$$N_\epsilon = \left\lceil \frac{2\ln(\epsilon/C_2)}{\ln\left(1 - \frac{\gamma^{\min}}{2\lambda^{\max}\|AA^\top\|}\right)} \right\rceil.$$

Proof. We will start with (i). After some algebra it can be shown that $\Psi_\lambda(v) = \frac{\lambda}{2}\|A^\top v\|^2 - \langle A^\top v, \tilde{y}_k\rangle + \omega^\star(v)$. From then on it's direct, and it has $\Psi_{\lambda_1}(v) - \Psi_{\lambda_2}(v) = \frac{\lambda_1-\lambda_2}{2}\|A^\top v\|^2 \geq 0$.

To see (ii), observe that $\Phi_\lambda^{(k)}$ is $L_\Phi^{\lambda,(k)}$ Lipschitz continuous from Assumption 3.6, 4.1. By its definition in (4.1), $\forall k \geq 0$ it satisfies $L_\Phi^{\lambda,(k)}$ is the largest when $\lambda = \lambda^{\max}$. And since $\lambda^{(k)}$ is finite, it has $L_\Phi^{\max} = \sup_{k\in\mathbb{N}\cup\{0\}} L_\Phi^{\lambda,(k)} < \infty$ too.

To see (iii), we use (i), (ii) and, $C_0 = \sup_{k\in\{0\}\cup\mathbb{N}} \left\|v_0^{(k)} - \bar{v}^{(k)}\right\| < \infty$ from Assumption 4.1, then it's sufficient to have: $C_1 = L_\Phi^{\max} C_0$ by $L_\Phi^{(k)}$ Lipschitz continuity of $\Phi_\lambda^{(k)}$.

To see (iv), use all previous results (iii), (ii) and, then it's sufficient to have $C_2 = \sqrt{C_1}\left(L_\Phi^{\max}\sqrt{2\lambda^{\max}} + \sqrt{C_1}\right)$.

And finally, we are ready to show (v), it has for all $k \ge 0$:

$$
\begin{aligned}
&\mathbf{G}_\lambda^{(k)}(z_j^{(k)}, v_j^{(k)})\\
&\underset{(1)}{\le} \left(1 - \frac{\gamma_\lambda^{(k)}}{2\tau^{(k)}}\right)^{j/2} \left(L_\Phi^{\lambda,(k)}\sqrt{2\lambda^{(k)}} + \sqrt{\Psi_\lambda^{(k)}(v_0^{(k)}) - \Psi_\lambda^{(k)}(\bar{v}^{(k)})}\right)\left(\Psi_\lambda^{(k)}(v_0^{(k)}) - \Psi_\lambda^{(k)}(\bar{v}^{(k)})\right)^{1/2}\\
&\underset{(2)}{\le} \left(1 - \frac{\gamma_\lambda^{(k)}}{2\tau^{(k)}}\right)^{j/2} \left(L_\Phi^{\lambda,(k)}\sqrt{2\lambda^{(k)}} + \sqrt{C_1}\right)\sqrt{C_1}\\
&\underset{(3)}{\le} \left(1 - \frac{\gamma_\lambda^{(k)}}{2\tau^{(k)}}\right)^{j/2} C_2\\
&\underset{(4)}{\le} \left(1 - \frac{\gamma^{\min}}{2\lambda^{\max}\|AA^\top\|}\right)^{j/2} C_2.
\end{aligned}
$$

At (1), we used Theorem 3.9. At (2), we used results form (iii) and, at (3) we used (iv). Finally, at (4), we use the fact that the step size τ in Definition 3.7 which translate into our context, it has $\tau^{(k)} \ge \lambda^{(k)}\|AA^\top\|$, and hence it can be upper bounded by $\tau^{\max} := \sup_{k\in\mathbb{N}\cup\{0\}} \lambda^{(k)}\|AA^\top\| < \infty$. Now, substitute $j = N_\epsilon$ then it's prominent that $\mathbf{G}_\lambda^{(k)}\left(z_j^{(k)}, v_j^{(k)}\right) \le \epsilon$ by using the log of base formula. ■

4.2 Overall complexity

In this section we combine the complexity of the inner loop with the convergence rate of the outer loop to deduce the rate of convergence. Details will be given to the assumptions needed to obtain globally bounded complexity of the inner loop.

A Super boring chores

Super boring stuff that I just want to skip but somehow necessary.

{lemma:chore1} **Lemma A.1 (That conjugate for the dual of proximal problem)**
Let $f : \mathbb{R}^n \to \overline{\mathbb{R}} := \frac{1}{2\lambda}\|u - v\|^2$, then its conjugate is given by

$$
f^\star(v) = \frac{1}{2\lambda}\|\lambda v + y\|^2 - \frac{1}{2\lambda}\|y\|^2.
$$

Proof. Recall the following properties for any closed, proper convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$. Let $a \in \mathbb{R}^n$ be any vector, let $\alpha \in \mathbb{R}, c \in \mathbb{R}$ be some scalar, then we have these three properties of conjugating convex funtion.

(i) $(\alpha f)^\star = \alpha f^\star \circ (\alpha^{-1} I)$.
(ii) $(f + c)^\star(y) = f^\star(y) - c$.
(iii) $(x \mapsto f(x) + \langle x, y\rangle)^\star (y) = f^\star(y - a)$.

From here we have:

$$\begin{aligned}
f^\star(v) &= \left(u \mapsto \lambda^{-1}\left(\frac{1}{2}\|u\|^2 - \langle u, y\rangle\right) + \frac{1}{2\lambda}\|y\|^2\right)^\star (v) \\
&= \left(u \mapsto \lambda^{-1}\left(\frac{1}{2}\|u\|^2 - \langle u, y\rangle\right)\right)^\star (v) - \frac{1}{2\lambda}\|y\|^2 \\
&= \left[\lambda^{-1}\left(u \mapsto \left(\frac{1}{2}\|u\|^2 - \langle u, y\rangle\right)\right)^\star \circ (\lambda I)\right](v) - \frac{1}{2\lambda}\|y\|^2 \\
&= \left[\lambda^{-1}\left(u \mapsto \left(\frac{\|\cdot\|^2}{2}\right)^\star (u + y)\right) \circ (\lambda I)\right](v) - \frac{1}{2\lambda}\|y\|^2 \\
&= \left[\lambda^{-1}\left(u \mapsto \frac{\|u + y\|^2}{2}\right) \circ (\lambda I)\right](v) - \frac{1}{2\lambda}\|y\|^2 \\
&= \lambda^{-1}\left(\frac{1}{2}\|\lambda v + y\|^2\right) - \frac{1}{2\lambda}\|y\|^2.
\end{aligned}$$

■

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.

[2] 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.

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

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

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

[6] 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.