# Reading Notes

Alto

Last Compiled: June 17, 2025

**Abstract**

Reports on papers read. This is a LaTEX file for my own notes taking. It may accelerate the process of writing my thesis for my PhD degree.

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

# Chapter 1

# The Basics of Optimization Theories

{def:bregman-div} Notations in this chapter are not shared, and they are for this chapter only.

**Definition 1.0.1 (Bregman Divergence)** *Let $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be a differentiable function. Define Bregman Divergence:*

{ass:smooth-add-nonsmooth}

$$D_f : \mathbb{R}^n \times \operatorname{dom} \nabla f \rightarrow \overline{\mathbb{R}} := (x, y) \mapsto f(x) - f(y) - \langle \nabla f(y), x - y \rangle.$$

**Assumption 1.0.2 (smooth plus nonsmooth)** Let $F = f + g$ where $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is differentiable and there exists $q \in \mathbb{R}$ such that $g - \mu/2\|\cdot\|^2$ is convex.

**Definition 1.0.3 (proximal gradient operator)** *Suppose $F = f + g$ satisfies Assumption 1.0.2. Let $\beta > 0$, we define the proximal gradient operator for all $x \in \mathbb{R}^n$:*

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

{thm:pg-ineq-swcnvx-generic} **Theorem 1.0.4 (strongly/weakly convex generic proximal gradient inequality)** *Suppose $F = f + g$ satisfies Assumption 1.0.2 with $\beta > 0$ and $\mu \in \mathbb{R}$. Then for all $x \in \mathbb{R}^n, z \in \mathbb{R}^n$, define $\bar{x} = T_{\beta^{-1}, f, g}(x)$, it has:*

$$\frac{\mu}{2}\|z - \bar{x}\|^2 \leq F(z) - F(\bar{x}) - \langle \beta(x - \bar{x}), z - \bar{x} \rangle + D_f(x, \bar{x}) - D_f(z, x).$$

*Proof.* Nonsmooth analysis calculus rules has

$$\begin{aligned}
&\bar{x} \in \operatorname{argmin} z \left\{ g(z) + \langle \nabla f(x), z \rangle + \frac{\beta}{2}\|z - x\|^2 \right\} \\
\implies &\mathbf{0} \in \partial g(\bar{x}) + \nabla f(x) + \beta(\bar{x} - x) \\
\iff &\partial g(x^+) \ni -\nabla f(x) - \beta(\bar{x} - x).
\end{aligned}$$

The subgradient inequality for weak convexity has

$$\begin{aligned}
\frac{\mu}{2}\|z-\bar{x}\|^2 &\le g(z) - g(\bar{x}) + \langle \nabla f(x) + \beta(\bar{x}-x), z-\bar{x}\rangle \\
&= g(z) - g(\bar{x}) + \langle \nabla f(x), z-\bar{x}\rangle + \langle \beta(\bar{x}-x), z-\bar{x}\rangle \\
&= g(z) - g(\bar{x}) + \langle \nabla f(x), z-x\rangle + \langle \nabla f(x), x-\bar{x}\rangle + \langle \beta(\bar{x}-x), z-\bar{x}\rangle \\
&= g(z) - g(\bar{x}) + (-D_f(z,x) + f(z) - f(x)) \\
&\quad + (D_f(\bar{x},x) - f(\bar{x}) + f(x)) + \langle \beta(\bar{x}-x), z-\bar{x}\rangle \\
&= F(z) - F(\bar{x}) - D_f(z,x) + D_f(\bar{x},x) - \langle \beta(x-\bar{x}), z-\bar{x}\rangle.
\end{aligned}$$

■

## 1.1 Convexity results

{thm:cnvx-pg-ineq}

**Theorem 1.1.1 (convex proximal gradient inequality)** *Suppose $F = f + g$ satisfies Assumption 1.0.2 such that $\mu = \mu_g \ge 0$, $\beta \ge L_f$. In addition, suppose that $f: \mathbb{R}^n \to \mathbb{R}$ has $L_f$ Lipschitz continuous gradient, and it's $\mu_f \ge 0$ strongly convex. For all $x \in \mathbb{R}^n, z \in \mathbb{R}^n$, define $\bar{x} = T_{\beta^{-1}, f, g}(x)$ it has*

$$0 \le F(z) - F(\bar{x}) + \frac{\beta - \mu_f}{2}\|z-x\|^2 - \frac{\beta + \mu_g}{2}\|z-\bar{x}\|^2.$$

*Proof.* The Bregman Divergence of $f$ has inequality

$$(\forall x \in \mathbb{R}^n, y \in \mathbb{R}^n)\ \frac{\mu_f}{2}\|x-y\|^2 \le D_f(x,y) \le \frac{L_f}{2}\|x-y\|^2.$$

Specializing Theorem 1.0.4, let $x \in \mathbb{R}^n$ and define $\bar{x} = T_{\beta^{-1}, f, g}(x)$ it has $\forall z \in \mathbb{R}^n$ :

$$\begin{aligned}
\frac{\mu_g}{2}\|z-\bar{x}\|^2 &\le F(z) - F(\bar{x}) - D_f(z,x) + D_f(\bar{x},x) - \langle \beta(x-\bar{x}), z-\bar{x}\rangle \\
&\le F(z) - F(\bar{x}) - \frac{\mu_f}{2}\|z-x\|^2 + \frac{L_f}{2}\|x-\bar{x}\|^2 - \langle \beta(x-\bar{x}), z-x+x-\bar{x}\rangle \\
&= F(z) - F(\bar{x}) - \frac{\mu_f}{2}\|z-x\|^2 + \left(\frac{L_f}{2} - \beta\right)\|x-\bar{x}\|^2 - \langle \beta(x-\bar{x}), z-x\rangle \\
&\le F(z) - F(\bar{x}) - \frac{\mu_f}{2}\|z-x\|^2 - \frac{\beta}{2}\|x-\bar{x}\|^2 - \langle \beta(x-\bar{x}), z-x\rangle \\
&= F(z) - F(\bar{x}) - \frac{\mu_f}{2}\|z-x\|^2 - \frac{\beta}{2}\left(\|x-\bar{x}\|^2 + 2\langle x-\bar{x}, z-x\rangle\right) \\
&= F(z) - F(\bar{x}) + \frac{\beta - \mu_f}{2}\|z-x\|^2 - \frac{\beta}{2}\|z-\bar{x}\|^2.
\end{aligned}$$

■

{def:scnvx}

**Definition 1.1.2 (strong convexity)** *Let $\alpha \geq 0$. $f : \mathbb{R}^n \to \mathbb{R}^n$ is $\alpha$ strongly convex if and only if $f - \alpha/2\|\cdot\|^2$ is a convex function.*

{thm:scnvx-eqvs} **Remark 1.1.3** The definition doesn't exclude convex functions.

**Theorem 1.1.4 (strong convexity equivalences)** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be $\alpha \geq 0$ strongly convex then, it is equivalent to*

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

# Chapter 2

# Linear Convergence of First Order Method

In this chapter, we are specifically interested in characterizing linear convergence of well known first order optimization algorithms. In this section, $D_f$ will denote the Bregman Divergence as defined in Definition 1.0.1.

Defined special symbols/substitutes in this chapter: Q-SCNVX $q\mathcal{S}$; QUA $\mathcal{U}$; QGG $\mathcal{G}$; QFG $\mathcal{F}$; PEB $\mathcal{E}$.

## 2.1 Necoara's et al.'s Paper

### 2.1.1 The Settings

{ass:necoara-2019-settings} The assumption follows give the same setting as Necoara et al. [9].

**Assumption 2.1.1** Consider optimization problem:

$$-\infty < f^+ = \min_{x \in X} f(x). \tag{2.1.1}$$

{problem:necoara-2019} $X \subseteq \mathbb{R}^n$ is a closed convex set. Assume projection onto $X$, denoted by $\Pi_X$ is easy. Denote $X^+ = \operatorname*{argmin}_{x \in X} f(x) \neq \emptyset$, assume it's a closed set. Assume $f$ has $L_f$ Lipschitz continuous gradient, i.e: for all $x, y \in X$:

$$\|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\|.$$

Some immediate consequences of Assumption 2.1.1 now follows. The variational inequality characterizing optimal solution has:

{ineq:pg-opt-cond}

$$x^+ \in X^+ \implies (\forall x \in X)\ \langle \nabla f(x^+), x - x^+ \rangle \geq 0. \tag{2.1.2}$$

The converse is true if $f$ is convex. We define the gradient mapping as:

$$\mathcal{G}_{L_f}(x) := L_f(x - \Pi_X(x - L_f^{-1}\nabla f(x))).$$

{thm:breg-div-scnvx}

**Theorem 2.1.2 (strong convexity characterized via Bregman Divergence)** *Suppose $f$ satisfies Assumption 2.1.1. Then $f$ is $\kappa_f$ strongly convex and $L_f$ smooth if*

$$(\forall x, y \in X)\ \kappa_f\|x - y\|^2 \leq D_f(x, y) \leq L_f\|x - y\|^2.$$

*We denote it by $f \in \mathcal{S}(L_f, \kappa_f, X)$.*

Then, it's not hard to imagine the following natural relaxations of the above condition.

**Definition 2.1.3 (relaxations of strong convexity)**
{def:necoara-weaker-scnvx} *Suppose $f$ satisfies Assumption 2.1.1. Let $L_f \geq \kappa_f \geq 0$ such that for all $x \in X$, $\bar{x} = \Pi_{X^+}x$.*
{def:neocara-qscnvx} *We define the following:*

(i) *Quasi-strong convexity (Q-SCNVX): $0 \leq D_f(\bar{x}, x) - \frac{\kappa_f}{2}\|x - \bar{x}\|^2$. Denoted by $q\mathcal{S}(L_f, \kappa_f, X)$.*

{def:necoara-qup}

(ii) *Quadratic under approximation (QUA): $0 \leq D_f(x, \bar{x}) - \frac{\kappa_f}{2}\|x - \bar{x}\|^2$. Denoted by $\mathcal{U}(L_f, \kappa_f, X)$.*

{def:necoara-qgg}

(iii) *Quadratic Gradient Growth (QGG): $0 \leq D_f(x, \bar{x}) + D_f(\bar{x}, x) - \kappa_f/2\|x - \bar{x}\|^2$. Denoted by $\mathcal{G}(L_f, \kappa_f, X)$.*

{def:necoara-qfg}

(iv) *Quadratic Function Growth (QFG): $0 \leq f(x) - f^* - \kappa_f/2\|x - \bar{x}\|^2$. Denoted by $\mathcal{F}(L_f, \kappa_f, X)$.*

{def:necoara-peb}

(v) *Proximal Error Bound (PEB): $\|\mathcal{G}_{L_f}x\| \geq \kappa_f\|x - \bar{x}\|$. Denoted by $\mathcal{E}(L_f, \kappa_f, X)$.*

**Remark 2.1.4** The error bound condition in Necoara et al. is sometimes referred to as the "Proximal Error Bound".

### 2.1.2 Weaker conditions of strong convexity

In Necoara's et al., major results assume convexity of $f$.

{thm:qscnvx-means-qua}

**Theorem 2.1.5 (Q-SCNVX implies QUA)** *Let $f$ satisfies Assumption 2.1.1 and assume $f$ is convex:*

$$q\mathcal{S}(L_f, \kappa_f, X) \subseteq \mathcal{U}(L_f, \kappa_f, X).$$

*Proof.* We prove by induction. Convexity of $f$ makes $X^+$ convex, so $\Pi_{X^+}x$ is unique for all $x \in \mathbb{R}^n$. Make inductive hypothesis that there exists $\kappa^{(k)} \geq 0$ such that

{ineq:thm:qscnvx-means-qua-proof-ih}

$$(\forall x \in X) \quad f(x) \geq f^+ + \langle \nabla f(\Pi_{X^+}x), x - \Pi_{X^+}x \rangle + \kappa_f^{(k)}/2\|x - \Pi_{X^+}x\|^2. \tag{2.1.3}$$

The base case is true by convexity of $f$ with $\kappa_f^{(0)} = 0$. Choose any $x \in X$ define $\bar{x} = \Pi_{X^+}x$. Consider $x_\tau = \bar{x} + \tau(x - \bar{x})$ for $\tau \in [0, 1]$. $f$ is Q-SCNVX so

{ineq:thm:qscnvx-means-qua-proof-item1}

$$\begin{aligned} f^+ - f(x_\tau) &\geq \langle \nabla f(x_\tau), \Pi_{X^+}x_\tau - x_\tau \rangle + \kappa_f/2\|x_\tau - \Pi_{X^+}x_\tau\|^2 \\ &= \langle \nabla f(x_\tau), \bar{x} - x_\tau \rangle + \kappa_f/2\|x_\tau - \bar{x}\|^2 \\ \iff \langle \nabla f(x_\tau), x_\tau - \bar{x} \rangle &\geq f(x_\tau) - f^+ + \kappa_f/2\|x_\tau - \bar{x}\|^2. \end{aligned} \tag{2.1.4}$$

In the inductive proof that comes, we will use the following intermediate results. They are labeled for ease of referencing.

(a) The inequality (2.1.4) we just showed.

(b) By the property of projection, it has $\Pi_{X^+}x_\tau = \bar{x}$.

(c) The inductive hypothesis, inequality (2.1.3).

(d) $\bar{x} = \Pi_{X^+}x$, $X^+$ is the set of minimizer of the of $f$ over $X$, hence $f(\bar{x}) = f^+$, the minimum.

Using calculus rules, we start with:

$$
\begin{aligned}
f(x) &= f(\bar{x}) + \int_0^1 \langle \nabla f(x_\tau), x - \bar{x}\rangle d\tau = f(\bar{x}) + \int_0^1 \tau^{-1}\langle \nabla f(x_\tau), \tau(x-\bar{x})\rangle d\tau \\
&= f(\bar{x}) + \int_0^1 \tau^{-1}\langle \nabla f(x_\tau), x_\tau - \bar{x}\rangle d\tau. \\
&\underset{(a)}{\geq} f(\bar{x}) + \int_0^1 \tau^{-1}\left(f(x_\tau) - f^+ + \frac{\kappa_f}{2}\|x_\tau - \bar{x}\|^2\right) d\tau = f(\bar{x}) + \int_0^1 \tau^{-1}\left(f(x_\tau) - f^+\right) + \frac{\tau\kappa_f}{2}\|x-\bar{x}\|^2 d\tau \\
&\underset{(c)}{\geq} f(\bar{x}) + \int_0^1 \tau^{-1}\left(\langle \nabla f(\Pi_{X^+}x_\tau), x_\tau - \Pi_{X^+}x_\tau\rangle + \frac{\kappa_f^{(k)}}{2}\|x_\tau - \Pi_{X^+}x_\tau\|^2\right) + \frac{\tau\kappa_f}{2}\|x - \Pi_{X^+}x_\tau\|^2 d\tau \\
&\underset{(b)}{=} f(\bar{x}) + \int_0^1 \tau^{-1}\left(\langle \nabla f(\bar{x}), x_\tau - \bar{x}\rangle + \frac{\kappa_f^{(k)}}{2}\|x_\tau - \bar{x}\|^2\right) + \frac{\tau\kappa_f}{2}\|x-\bar{x}\|^2 d\tau \\
&= f(\bar{x}) + \int_0^1 \langle \nabla f(\bar{x}), x - \bar{x}\rangle + \frac{\tau\kappa_f^{(k)}}{2}\|x-\bar{x}\|^2 + \frac{\tau\kappa_f}{2}\|x-\bar{x}\|^2 d\tau \\
&\underset{(d)}{=} f^+ + \langle \nabla f(\bar{x}), x - \bar{x}\rangle + \frac{\kappa_f^{(k)} + \kappa_f}{4}\|x-\bar{x}\|^2.
\end{aligned}
$$

This is the new inductive hypothesis, and it has $\kappa_f^{(k+1)} = (\kappa_f^{(k)} + \kappa_f)/2$. The induction admits recurrence:

$$
\kappa_f^{(n)} = (1/2^n)(\kappa_f^{(0)} + (2^n - 1)\kappa_f).
$$

Inductive hypothesis is true for $\kappa_f^{(0)} = 0$ and $f$ being convex is sufficient. It has $\lim_{n\to\infty} \kappa_f^{(n)} = \kappa_f$. ■

**Remark 2.1.6** This is Theorem 1 in the paper. Convexity assumption of $f$ makes $X^+$ convex, so the projection is unique, and it has $\Pi_{X^+}x_\tau = \bar{x}$ for all $\tau \in [0,1]$. In addition, the inductive hypothesis has $\kappa_f^{(n)} \geq 0$, which is not sufficient for convexity, but necessary. The projection property remains true for nonconvex $X^+$, however the base case require rethinking.

{thm:qgg-implies-qua}

**Theorem 2.1.7 (QGG implies QUA)** *Let $f$ satisfies Assumption 2.1.1, under convexity it has*

$$
\mathcal{G}(L_f, \kappa_f, X) \subseteq \mathcal{U}(L_f, \kappa_f, X).
$$

*Proof.* For all $x \in X$, define $\bar{x} = \Pi_{X^+}x$, $x_\tau = \bar{x} + \tau(x - \bar{x})\ \forall \tau \in [0, 1]$. Observe that $\frac{d}{d\tau}x_\tau = x - \bar{x}$ and $\Pi_{X^+}x_\tau = \bar{x}\ \forall \tau \in [0, 1]$. Using calculus, Definition 2.1.3 (iii):

$$
\begin{aligned}
f(x) &= f(\bar{x}) + \int_0^1 \langle \nabla f(x_\tau), x - \bar{x}\rangle d\tau \\
&= f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle + \int_0^1 \langle \nabla f(x_\tau) - \nabla f(\bar{x}), x - \bar{x}\rangle d\tau \\
&= f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle + \int_0^1 \tau^{-1}\langle \nabla f(x_\tau) - \nabla f(\bar{x}), \tau(x - \bar{x})\rangle d\tau \\
&= f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle + \int_0^1 \tau^{-1}\langle \nabla f(x_\tau) - \nabla f(\bar{x}), x_\tau - \bar{x}\rangle d\tau \\
&\ge f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle + \int_0^1 \tau^{-1}\kappa_f\|\tau(x - \bar{x})\|^2 d\tau \\
&= f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle + \int_0^1 \tau\kappa_f\|x - \bar{x}\|^2 d\tau \\
&= f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle + \frac{\kappa}{2}\|x - \bar{x}\|^2.
\end{aligned}
$$

■

**Remark 2.1.8** This is Theorem 3 in Neocara et al. [9]. There is no immediate use of convexity besides that the projection $\bar{x} = \Pi_{X^+}x$ is a singleton.

{thm:qscnvx-implies-qgg}

**Theorem 2.1.9 (Q-SCNVX implies QGG)** *Under Assumption 2.1.1 and convexity of $f$, it has*

$$
q\mathcal{S}(L_f, \kappa_f, X) \subseteq \mathcal{G}(L_f, \kappa_f, X).
$$

*Proof.* If $f \in q\mathcal{S}(L_f, \kappa_f, X)$ then Theorem 2.1.5 has $f \in \mathcal{U}(L_f, \kappa_f, X)$. Then, add (ii), (i) in Definition 2.1.3 yield the results. ■

**Remark 2.1.10** This is Theorem 2 in the Necoara et al. [9], right after it claims $\mathcal{U}(L_f, \kappa_f, X) \subseteq \mathcal{G}(L_f, \kappa_f/2, X)$ under convexity.

Prior to proving more theorems, we will specialize Theorem 1.1.1 because it will be used in later proofs. In Assumption 2.1.1, it can be seemed as taking $F = f + g$ in Assumption 1.0.2 with $g = \delta_X$. This makes $\mu_g = 0$ and assuming $f$ is convex we have $\mu_f = 0$. Let $\beta = L_f$, and $x^+ = \Pi_X(x - L_f^{-1}\nabla f(x))$, it has for all $z \in X$:

{ineq:proj-grad}

$$
\begin{aligned}
0 &\le f(z) - f(x^+) + \frac{L_f}{2}\|z - x\|^2 - \frac{L_f}{2}\|z - x^+\|^2 \\
&= f(z) - f(x^+) + L_f\langle z - x^+, x^+ - x\rangle + \frac{L_f}{2}\|x - x^+\|^2.
\end{aligned}
\tag{2.1.5}
$$

Take note that when $z = x$ it has

{ineq:proj-grad2}

$$0 \le f(x) - f(x^+) - \frac{L_f}{2}\|x - x^+\|^2. \tag{2.1.6}$$

{thm:qfg-suff}

**Theorem 2.1.11 (sufficiency of QFG)** *Let $f$ satisfies Assumption 2.1.1. For all $0 < \beta < 1$, $x \in X$, let $x^+ = \Pi_X(x - L_f^{-1}\nabla f(x))$. If*

$$\|x^+ - \Pi_{X^+}x^+\| \le \beta\|x - \Pi_{X^+}x\|,$$

*then $f$ satisfies the QFG condition with $\kappa_f = L_f(1-\beta)^2$.*

*Proof.* The proof is direct. The first inequality is the property of projection onto a set.

$$\begin{aligned}
\|x - \Pi_{X^+}x\| &\le \|x - \Pi_{X^+}x^+\| \\
&\le \|x - x^+\| + \|x^+ - \Pi_{X^+}x^+\| \\
&\le \|x - x^+\| + \beta\|x - \Pi_{X^+}x\| \\
\iff 0 &\le \|x - x^+\| - (1-\beta)\|x - \Pi_{X^+}x\|.
\end{aligned}$$

Now, use (2.1.6).

$$f(x^+) - f(x) \le -\frac{L_f}{2}\|x^+ - x\|^2 \le -\frac{L_f}{2}(1-\beta)^2\|x - \Pi_{X^+}x\|^2.$$

Hence, it gives the quadratic growth condition. ∎

**Remark 2.1.12** It's unclear where convexity is used. Take the careful observation that inequality (2.1.6) doesn't still hold for weakly convex function. However, it' still assumed in Necoara et al. paper.

The following theorems are about the relation between PEB and QFG.

{lemma:grad-map-qfg}

**Lemma 2.1.13 (gradient mapping and quadratic function growth)** *Let $f$ satisfies Assumption 2.1.1. Suppose that $f \in \mathcal{F}(L_f, \mu_f, X)$ so, it satisfies the quadratic function growth condition. For all $x \in \mathbb{R}^n$, define $x^+ = \Pi_X(x - L_f^{-1}\nabla f(x))$, define projections onto the set of minimizers $\bar{x}^+ = \Pi_{X^+}x^+, \bar{x} = \Pi_{X^+}x$, then*

$$\left(\sqrt{L_f(\kappa_f + L_f)} - L_f\right)\|x^+ - \bar{x}^+\| \le \|L_f(x - x^+)\|.$$

*Proof.* Using convexity, consider projected gradient inequality (2.1.5) with $z = \bar{x}^+$ it yields:

$$
\begin{aligned}
0 &\geq f(x^+) - f(\bar{x}^+) - L_f\langle \bar{x}^+ - x^+, x^+ - x\rangle - \frac{1}{2L_f}\|L_f(x - x^+)\|^2 \\
&\underset{(a)}{\geq} \frac{\kappa_f}{2}\|x^+ - \bar{x}^+\|^2 - \|L_f(x - x^+)\|\|\bar{x}^+ - x^+\| - \frac{1}{2L_f}\|L_f(x - x^+)\|^2 \\
&= \frac{\kappa_f}{2}\|x^+ - \bar{x}^+\|^2 - \frac{1}{2L_f}\left(\|L_f(x - x^+)\|^2 + 2L_f\|L_f(x - x^+)\|\|\bar{x}^+ - x^+\|\right) \\
&= \frac{\kappa_f + L_f}{2}\|x^+ - \bar{x}^+\|^2 - \frac{1}{2L_f}\left(\|L_f(x - x^+)\|^2 + 2L_f\|L_f(x - x^+)\|\|\bar{x}^+ - x^+\| + L_f^2\|\bar{x}^+ - x^+\|^2\right) \\
&= \frac{\kappa_f + L_f}{2}\|x^+ - \bar{x}^+\|^2 - \frac{1}{2L_f}\left(\|L_f(x - x^+)\| + L_f\|x - \bar{x}^+\|\right)^2.
\end{aligned}
$$

At label (a), we statement hypothesis $f \in \mathcal{F}(L_f, \kappa_f, X)$ and, Cauchy inequality. From the last line, it can be equivalently expressed as:

$$
\begin{aligned}
0 &\leq \|L_f(x - x^+)\| + L_f\|x^+ - \bar{x}^+\| - \sqrt{L_f(\kappa_f + L_f)}\|x^+ - \bar{x}^+\| \\
&= \|L_f(x - x^+)\| - \left(\sqrt{L_f(\kappa_f + L_f)} - L_f\right)\|x^+ - \bar{x}^+\|.
\end{aligned}
$$

■

{thm:qfg-peb-equiv}

**Theorem 2.1.14 (equivalence between QFG and PEB)** *If $f$ is convex and satisfies Assumption 2.1.1. Then we have:*

$$
\begin{aligned}
\mathcal{F}(L_f, \kappa_f) &\subseteq \mathcal{E}\left(L_f, \frac{\kappa_f}{\kappa_f/L_f + 1 + \sqrt{\kappa_k/L_f + 1}}, X\right), \\
\mathcal{E}(L_f, \kappa_f, X) &\subseteq \mathcal{F}(L_f, \kappa_f^2/L_f, X).
\end{aligned}
$$

*Proof.* For any $x \in X$, define the projected gradient: $x^+ = \Pi_X(x - L_f^{-1}\nabla f(x))$. Denote $\bar{x}^+ = \Pi_{X^+}x^+$. Let $\bar{x} = \Pi_{X^+}x$, using the property of projection onto $X$ we have

{ineq:thm:qfg-peb-equiv-proof-item1}

$$
\begin{aligned}
\|x - \bar{x}\| &\leq \|x - \bar{x}^+\| \leq \|x - x^+\| + \|x^+ - \bar{x}^+\| \\
&= \frac{1}{L_f}\|L_f(x - x^+)\| + \|x^+ - \bar{x}^+\| \\
\iff \|x^+ - \bar{x}^+\| &\geq \|x - \bar{x}\| - \frac{1}{L_f}\|L_f(x - x^+)\|. \qquad (2.1.7)
\end{aligned}
$$

Starting with Lemma 2.1.13 because $f$ satisfies quadratic growth and, it is assumed convex, then it has:

$$
\begin{aligned}
0 &\le \|L_f(x - x^+)\| - \left(\sqrt{L_f(\kappa_f + L_f)} - L_f\right)\|x^+ - \bar{x}^+\| \\
&\underset{\text{(a)}}{\le} \|L_f(x - x^+)\| - \left(\sqrt{L_f(\kappa_f + L_f)} - L_f\right)\left(\|x - \bar{x}\| - \frac{1}{L_f}\|L_f(x - x^+)\|\right) \\
&= -\left(\sqrt{L_f(\kappa_f + L_f)} - L_f\right)\|x - \bar{x}\| + \left(L_f^{-1}\left(\sqrt{L_f(\kappa_f + L_f)} - L_f\right) + 1\right)\|L_f(x - x^+)\| \\
&= -\left(\sqrt{L_f(\kappa_f + L_f)} - L_f\right)\|x - \bar{x}\| + \sqrt{L_f(\kappa_f + L_f)}\|L_f(x - x^+)\| \\
\iff & \frac{\sqrt{L_f(\kappa_f + L_f)} - L_f}{\sqrt{L_f(\kappa_f + L_f)}}\|x - \bar{x}\| \le L_f^{-1}\|\mathcal{G}_{L_f}x\|.
\end{aligned}
$$

At the label (a), we used inequality (2.1.7). Continuing with some algebra, we have:

$$
\begin{aligned}
\frac{\sqrt{L_f(\kappa_f + L_f)} - L_f}{\sqrt{L_f(\kappa_f + L_f)}} &= \frac{L_f(\kappa_f + L_f) - L_f^2}{\sqrt{L_f(\kappa_f + L_f)}(\sqrt{L_f(\kappa_f + L_f)} + L_f)} \\
&= \frac{L_f\kappa_f}{L_f(\kappa_f + L_f) + L_f\sqrt{L_f(\kappa_f + L_f)}} \\
&= \frac{\kappa_f/L_f}{\kappa_f/L_f + 1 + \sqrt{\kappa_k/L_f + 1}}.
\end{aligned}
$$

This gives PEB condition:

$$
\|\mathcal{G}_{L_f}(x)\| \ge L_f\frac{\kappa_f/L_f}{\kappa_f/L_f + 1 + \sqrt{\kappa_k/L_f + 1}}\|x - \bar{x}\| = \frac{\kappa_f}{\kappa_f/L_f + 1 + \sqrt{\kappa_k/L_f + 1}}\|x - \bar{x}\|.
$$

**We now show PEB implies QFG**. From the error bound condition using $\kappa_f$ it has

$$
\kappa_f^2\|x - \bar{x}\|^2 \le \|\mathcal{G}_{L_f}(x)\|^2 \underset{(2.1.6)}{\le} 2L_f(f(x) - f(x^+)) \le 2L_f(f(x) - f^+).
$$

■

The following theorem summarizes the hierarchy of the conditions listed in Definition 2.1.3.

{thm:wcnvx-hierach}

**Theorem 2.1.15 (Hierarchy of weaker SCNVX conditions)** *Let $f$ satisfy Assumption 2.1.1, assuming convexity then the following relations are true:*

$$
\mathcal{S}(\kappa_f, L_f, X) \subseteq q\mathcal{S}(\kappa_f, L_f, X) \subseteq \mathcal{G}(\kappa_f, L_f, X) \subseteq \mathcal{U}(\kappa_f, L_f, X) \subseteq \mathcal{F}(\kappa_f, L_f, X).
$$

*Proof.* $q\mathcal{S} \subseteq \mathcal{G}$ is proved in Theorem 2.1.9 and $\mathcal{G} \subseteq \mathcal{U}$ is proved in Theorem 2.1.7. $\mathcal{S} \subseteq q\mathcal{S}'$ is obvious. It remains to show $\mathcal{U} \subseteq \mathcal{F}$. Let $f \in \mathcal{U}(\kappa_f, L_f, X)$, it has for all $x \in X$:

$$
\begin{aligned}
0 &\le f(x) - f^+ - \langle \nabla f(\bar{x}), x - \bar{x}\rangle - \frac{\kappa_f}{2}\|x - \bar{x}\|^2 \\
&\underset{(2.1.2)}{\le} f(x) - f^+ - \frac{\kappa_f}{2}\|x - \bar{x}\|^2.
\end{aligned}
$$

■

**Remark 2.1.16** It's Theorem 4 in Necoara et al. [9].

### 2.1.3 Characterizing Q-SCNVX functions with Hoffman's error bound

Back in 1955, Hoffman [8] determined the existence of an upper bound for the distance of a point to a polyhedron can be bounded by the constraint violations of the point, uniformly. Let's adopt $[\cdot]_+$ as the notation for projecting on $\mathbb{R}^n_+$. Necoara re-introduce the bound, the following theorem phrase that upper bound in the exact same way, but only with the Euclidean norm.

**Theorem 2.1.17 (Polyhedral Hoffman Error bound)** *Let $A \in \mathbb{R}^{p\times n}, C \in \mathbb{R}^{m\times n}$. Given non-empty polyhedron: $P := \{x \in \mathbb{R}^n : Ax = b, Cx \le d\}$, there exists a constant $\theta(A, C) > 0$ such that*

$$
(\forall x \in \mathbb{R}^n)\ \operatorname{dist}(x|P) \le \theta(A, C) \left\| \begin{bmatrix} Ax - b \\ [Cx - d]_+ \end{bmatrix} \right\|.
$$

The constant $\theta(A, C)$ is referred to as the Hoffman constant in the literature. There are two notable generalization of the inequality. Güler et al. [7] found a representation of the Hoffman constant via singular values of sub matrices that defines the polyhedron. Burke and Tseng [4] unified the analysis for the Hoffman constant and extend it to the context of general cone over a sublinear convex function, and more.

It's important to note at this point that the exact value of $\theta(A, C)$ is very difficult to obtain in practice. Let's go back to Necoara [9]. The following theorem stated their use of the this error bound to capture a class of Q-SCNVX function.

{thm:q-scnvx-hffmn-eb}

**Theorem 2.1.18 (Q-SCNVX and Hoffman's error bound)** *Let $X := \{S \in \mathbb{R}^n : Cx \le d\}$ be non-empty, $g : \mathbb{R}^m \to \mathbb{R}$ be a $\sigma_g$ strongly convex and, $L_g$ Lipschitz smooth function. Let $A \in \mathbb{R}^{m\times n}$ be a nonzero, $C \in \mathbb{R}^{m\times n}$. Define $f(x) := g(Ax)$, then $f$ is Q-SCNVX with $L_f = L_g\|A\|^2$, $\kappa_f = \sigma_g/\theta(A, C)^2$.*

*Proof.* $X \neq \emptyset$ gives a non-empty set of minimizers $X^+ := \operatorname*{argmin}_{x\in X} f(x)$. We state the following intermediate results prior to their justifications at the end of the proof.

(a) $\forall x, y$ it has $\|\nabla f(x) - \nabla f(y)\| \leq \|A\|^2 L_g \|x - y\|$.

(b) There exists uniquely $(t^*, T^*) \in \mathbb{R}^{m\times n}$ such that for all $\bar{x} \in X^+$ it has $A\bar{x} = t^*, \nabla f(\bar{x}) = T^*$. In other word, for all minimizers $\bar{x}$ in $X^+$, $A\bar{x}$ and, $\nabla f(\bar{x})$ are all the same. It would mean $X^+ = \{\bar{x} : A\bar{x} = t^*, C\bar{x} \leq d\}$ and, it's also a polyhedron.

(c) Hoffman error bound for Theorem 2.1.18 applied to polyhedron $X^+$, it has

$$\operatorname{dist}(x|\ X^+) \leq \theta(A, C) \operatorname{dist}((Ax - t^*, Cx - d)|\ \{\mathbf{0}\} \times \mathbb{R}^n_-)$$

And for any feasible solution $x \in X$, it simplifies to $\operatorname{dist}(x|X^+) \leq \theta(A, C) \operatorname{dist}(Ax - t^*|\ \{\mathbf{0}\} \times \mathbb{R}^n_-)$.

Snatching these two intermediate results, the proof for the overall theorem follows. For all $x \in X$ denote $\bar{x} = \Pi_{X^+}x$, using strong convexity of $g$, and $f(x) = g(Ax)$:

$$\begin{aligned}
f(\bar{x}) - f(x) - \langle \nabla f(x), \bar{x} - x\rangle &= g(A\bar{x}) - g(x) - \left\langle A^T\nabla g(Ax), \bar{x} - x\right\rangle \\
&= g(A\bar{x}) - g(x) - \langle \nabla g(Ax), A\bar{x} - Ax\rangle \\
&\geq \sigma_g/2\|Ax - A\bar{x}\|^2 \\
&\underset{\text{(b)}}{=} \frac{\sigma_g}{2}\|Ax - t^*\|^2 \\
&\underset{\text{(c)}}{\geq} \frac{\sigma_g}{2\theta(A, C)^2}\|x - \bar{x}\|^2.
\end{aligned}$$

Therefore, the function $f$ is Q-SCNVX, i.e: $f \in q\mathcal{S}(L_g\|A\|^2, \sigma_g/\theta(A, C)^2)$.
**The Proof of (a)** is direct.

$$\begin{aligned}
\|\nabla f(x) - \nabla f(y)\| &= \|A^T\nabla g(Ax) - A^T\nabla g(Ay)\| \\
&\leq \|A\|\|\nabla g(Ax) - \nabla g(Ax)\| \\
&\leq \|A\|L_g\|Ax - Ay\| \leq \|A\|^2 L_g\|x - y\|.
\end{aligned}$$

**The proof of (b)** has some details. The set $X^+$ is convex and, $f$ takes the same value all over it. Consider any $\bar{x}_1, \bar{x}_2 \in X^+$, they have

$$\begin{aligned}
(1/2)(f(\bar{x}_1) + f(\bar{x}_2)) &= f((1/2)(\bar{x}_1 + \bar{x}_2)) \\
&= g((1/2)(A\bar{x}_1 + \bar{x}_2)) \\
&= (1/2)(g(A\bar{x}_1) + g(A\bar{x}_2)) \\
&\leq (1/2)g(A\bar{x}_1) + (1/2)g(A\bar{x}_2) - \frac{\sigma_g}{8}\|A\bar{x}_1 + A\bar{x}_2\|^2.
\end{aligned}$$

On the last inequality, we used an equivalent characterization of strong convexity (Theorem 1.1.4). The last inequality implies that $0 \leq -\|A\bar{x}_1 - A\bar{x}_2\|^2$, hence it must be that $A\bar{x}_1 = A\bar{x}_2$, which would mean $\nabla f(\bar{x}_1) = \nabla f(\bar{x}_2)$ because $\nabla f(x) = A^T\nabla g(A\bar{x})$. ■

### 2.1.4 Feasible descent and accelerated feasible descent

This section summarizes results from Necoara et al. on the method of feasible descent, fast feasible descent, and fast feasible descent with restart.

{def:projg-alg}

**Definition 2.1.19 (projected gradient algorithm)**
*The projected gradient algorithm generates a sequence of iterates* $(x_k)_{k\geq 0}$ *such that they satisfy for all* $k \geq 0$

$$x_{k+1} = \Pi_X(x_k - \alpha_k \nabla f(x_k)),$$

*Where* $\alpha_k \geq L_f^{-1}$ *for all* $k \geq 1$.

Under Assumption 2.1.1, convexity of $X$ means obtuse angle theorem from projection, and it specializes to

{ineq:projg-variational-ineq}

$$(\forall x \in X)\ \langle x_{k+1} - (x_k + \alpha_k \nabla f(x_k)), x_{k+1} - x\rangle \leq 0. \tag{2.1.8}$$

**Theorem 2.1.20** *feasible descent linear convergence under Q-SCNVX Under Assumption 2.1.1, assume that* $f$ *is Q-SCNVX with* $\mu_f, L_f$*, then the sequence that satisfies Definition 2.1.19 has a linear convergence rate. Let* $\bar{x}_k = \Pi_{X^+}x_k, \bar{x}_0 = \Pi_{X^+}x_0$. *For all* $k \geq 1$*, the iterates satisfy*

$$\|x_k - \bar{x}_k\|^2 \leq \left(\frac{1 - \kappa_f/L_f}{1 + \kappa_f/L_f}\right)^k \|x_0 - \bar{x}_0\|^2.$$

*Proof.* Our proof makes use of the following properties which we label it in advance for swift exposition:

(i) Inequality (2.1.8), from the projected gradient and convexity of $X$.

(ii) $f \in \mathbb{S}'$ which is the hypothesis that $f$ is Q-CNVX.

(iii) $\alpha_k \leq L_f^{-1}$, the stepsize is sufficient to apply descent lemma globally.

(iv) $f \in \mathbb{Q}$ satisfying Q-Growth, a consequence of Q-SCNVX by Theorem 2.1.15.

With $\overline{(\cdot)} = \Pi_{X^+}(\cdot)$ to denote the projection of a vector to the set of minimizers. The sequence of inequalities and equalities proves the theorem.

$$\begin{aligned}
\|x_{k+1} - \bar{x}_k\|^2 &= \|x_{k+1} - x_k + x_k - \bar{x}_k\|^2 = \|x_{k+1} - x_k\|^2 + \|x_k - \bar{x}_k\|^2 + 2\langle x_{k+1} - x_k, x_k - \bar{x}_k\rangle \\
&= (-\|x_{k+1} - x_k\|^2 + \|x_k - \bar{x}_k\|^2) + 2\|x_{k+1} - x_k\|^2 + 2\langle x_{k+1} - x_k, x_k - \bar{x}_k\rangle \\
&= -\|x_{k+1} - x_k\|^2 + \|x_k - \bar{x}_k\|^2 + 2\langle x_{k+1} - x_k, x_{k+1} - \bar{x}_k\rangle \\
&= -\|x_{k+1} - x_k\|^2 + \|x_k - \bar{x}_k\|^2
\end{aligned}$$

$$
\begin{aligned}
&+ 2\langle x_{k+1} - x_k + \alpha_k \nabla f(x_k), x_{k+1} - \bar{x}_k \rangle - 2\alpha_k \langle \nabla f(x_k), x_{k+1} - \bar{x}_k \rangle \\
&\underset{\text{(i)}}{\leq} -\|x_{k+1} - x_k\|^2 + \|x_k - \bar{x}_k\|^2 - 2\alpha_k \langle \nabla f(x_k), x_{k+1} - \bar{x}_k \rangle \\
&= -\|x_{k+1} - x_k\|^2 + \|x_k - \bar{x}_k\|^2 + 2\alpha_k \langle \nabla f(x_k), \bar{x}_k - x_k \rangle + 2\alpha_k \langle \nabla f(x_k), x_k - x_{k+1} \rangle \\
&\underset{\text{(ii)}}{\leq} -\|x_{k+1} - x_k\|^2 + \|x_k - \bar{x}_k\|^2 \\
&\quad + 2\alpha_k \left( f^+ - f(x_k) - \frac{\kappa_f}{2} \|x_k - \bar{x}_k\|^2 \right) + 2\alpha_k \langle \nabla f(x_k), x_k - x_{k+1} \rangle \\
&= (1 - \alpha_k \kappa_f) \|x_k - \bar{x}_k\|^2 \\
&\quad + 2\alpha_k (f^+ - f(x_k)) - 2\alpha_k \left( \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{1}{2\alpha_k} \|x_{k+1} - x_k\|^2 \right) \\
&= (1 - \alpha_k \kappa_f) \|x_k - \bar{x}_k\|^2 + 2\alpha_k f^+ \\
&\quad - 2\alpha_k \left( f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{1}{2\alpha_k} \|x_{k+1} - x_k\|^2 \right) \\
&\underset{\text{(iii)}}{\leq} (1 - \alpha_k \kappa_f) \|x_k - \bar{x}_k\|^2 + 2\alpha_k f^+ \\
&\quad - 2\alpha_k \left( f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L_f}{2} \|x_{k+1} - x_k\|^2 \right) \\
&\leq (1 - \alpha_k \kappa_f) \|x_k - \bar{x}_k\|^2 + 2\alpha_k f^+ - 2\alpha_k f(x_{k+1}) \\
&\underset{\text{(iv)}}{\leq} (1 - \alpha_k \kappa_f) \|x_k - \bar{x}_k\|^2 - \alpha_k \kappa_k \|x_{k+1} - \bar{x}_{k+1}\|^2.
\end{aligned}
$$

Therefore, it has

$$
\begin{aligned}
0 &\leq \|x_{k+1} - \bar{x}_k\|^2 - \|x_{k+1} - \bar{x}_{k+1}\|^2 \\
&\leq (1 - \alpha_k \kappa_f) \|x_k - \bar{x}_k\|^2 - \alpha_k \kappa_k \|x_{k+1} - \bar{x}_{k+1}\|^2 - \|x_{k+1} - \bar{x}_{k+1}\|^2 \\
&= (1 - \alpha_k \kappa_f) \|x_k - \bar{x}_k\|^2 - (1 + \alpha_k \kappa_k) \|x_{k+1} - \bar{x}_{k+1}\|^2.
\end{aligned}
$$

Unrolling recursively, then use (iii), the claim is proved. ■

# Chapter 3

# Application, Linear Feasibility Problems

This chapter extends ideas by the end of Necoara et al.'s paper [9].

## 3.1 Reducing LP to linear feasibility problems

{def:lcf-problem} We first introduce the linear feasibility problem as our primary problem.

**Definition 3.1.1 (a linear conic feasibility problem)** *Let $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^{m\times p}$ Let $\mathcal{K} \in \mathbb{R}^{n\times p}$ nonempty, closed convex cone. The conic feasibility problem is defined as the following optimization problem:*

$$\min_{X\in\mathcal{K}} \|AX - B\|_F^2.$$

The KKT of a linear programming problem is an instance of a linear feasibility problem, with $\mathcal{K}$ being a cross product of one of $\mathbb{R}, \mathbb{R}_+, \mathbb{R}_-$ and $p = 1$ making $x \in \mathbb{R}^n$. A feasibility semidefinite program with linear constraint is another example.

### 3.1.1 Example, linear programming is a linear conic feasibility problem

Let $X_1, X_2, Y$ be Euclidean spaces. Define linear mapping $E : X_1 \times X_2 \to Y := (x_1, x_2) \mapsto E_1x_1 + E_2x_2$ where $E_1, E_2$ each are mappings of $X_1 \to Y, X_2 \to Y$. Denote the adjoint of linear mapping by $(\cdot)^*$. Let $c = (c_1, c_2) \in X_1 \times X_2$, $b \in Y$. Suppose that $\mathcal{K} \subseteq X_1$ is a simple cone and $K^*$ is its dual cone. We consider the following linear programming problem

{problem:lp-cannon-form}

$$\inf_{x \in X_1 \times X_2} \left\{\langle -c, x\rangle \mid Ex = b, x \in \mathcal{K} \times X_2\right\}. \tag{3.1.1}$$

Define linear mapping $g, F$ and indicator function $h$ by the following:

$$\begin{aligned}
g : X_1 \times X_2 \to \mathbb{R} &:= x \mapsto \langle -c, x\rangle, \\
F : X_1 \times X_2 \to Y \times X_1 &:= (x_1, x_2) \mapsto (E_1 x_1 + E_2 x_2, x_1), \\
h : Y \times X_1 \to \overline{\mathbb{R}} &:= (y, z) \mapsto \delta_{\{\mathbf{0}\}}(y - b) + \delta_{\mathcal{K}}(z).
\end{aligned}$$

It's not hard to identify that problem in (3.1.1) has representations

$$\inf_{x \in X_1 \times X_2} \left\{g(x) + h(Fx)\right\}.$$

The dual problem of the above is given by

$$-\inf_{u \in Y \times X_1} \left\{h^\star(u) + g^\star(-F^* u)\right\}.$$

Where $h^\star, g^\star$ are the conjugate of $h, g$ and $F^* : Y \times X_1 \to X_1 \times X_2 = (y, z) \mapsto (E_1^* y + z, E_2^* y)$ is the adjoint operator of $F$. Note that $g^\star(x) = \delta_{\mathbf{0}}(x + c)$ and $h^\star((y, z)) = \langle b, y\rangle + \delta_{\mathcal{K}^*}(z)$. This gives the following dual problem

$$-\inf_{(y,z) \in Y \times \mathcal{K}^*} \left\{\langle b, y\rangle \mid E_1^* y + z = c_1, E_2^* y = c_2\right\}.$$

The KKT conditions give the following linear feasibility problem

$$\begin{aligned}
E_1 x_1 + E_2 x_2 &= b, \\
E_1^* y + z &= c_1, \\
E_2^* y &= c_2, \\
\langle b, y\rangle &= \langle c_1, x_1\rangle + \langle c_2, x_2\rangle, \\
(x_1, x_2) &\in \mathcal{K} \times X_2, \\
(y, z) &\in Y \times \mathcal{K}^*.
\end{aligned}$$

Assuming $X_1 = \mathbb{R}^{n_1}, X_2 = \mathbb{R}^{n_2}, Y = \mathbb{R}^m$. Define

$$\begin{aligned}
&\mathbf{K} := \mathcal{K} \times \mathbb{R}^{n_2} \times \mathbb{R}^m \times \mathcal{K}^*, \\
&A := \begin{bmatrix} E_1 & E_2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & E_1^T & I_{n_1} \\ \mathbf{0} & \mathbf{0} & E_2^T & \mathbf{0} \\ c_1^T & c_2^T & -b^T & 0 \end{bmatrix}, v := \begin{bmatrix} x_1 \\ x_2 \\ y \\ z \end{bmatrix} \in \mathbf{K}, d := \begin{bmatrix} b \\ c_1 \\ c_2 \\ 0 \end{bmatrix}.
\end{aligned}$$

The KKT conditions is a convex feasibility problem which can be formulated by best approximation problem:

{problem:lp-kkt-min}

$$\min_{v\in\mathbf{K}} \frac{1}{2}\|Ax-d\|^2. \tag{3.1.2}$$

It is minimizing a quadratic problem on a simple cone. Solving (3.1.1) can be approached by optimizing (3.1.2). It's necessary to investigate the matrices $A, A^T$ which are essential to solving it numerically. The properties of $A^TA$ will determine the convergence rate of algorithms. The matrix is a block matrix and possibly sparse in practice. Let $v = (x_1, x_2, y, z)$, it admits implicit representation:

$$Av = (E_1x_1 + E_2x_2,\ E_1^Ty + z,\ E_2^Ty,\ c_1^Tx_1 + c_2^Tx_2 - b^Ty).$$

It involves

(i) Two multiplications of $E$: $x_1, x_2$ on the right and $y$ on the right,

(ii) inner product using $x_1, x_2$ and $y$.

Let $\bar v = (\bar y, \bar x_1, \bar x_2, \xi) \in \mathbb{R}^m \times \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \mathbb{R}$ then the right multiplication of has:

$$\begin{aligned}\bar v^TA &= (E_1^T\bar y + \xi c_1^T,\ E_2^T\bar y + \xi c_2^T,\ \bar x_1^TE_1^T + \bar x_2^TE_2^T - \xi b^T,\ \bar x_1^T)\\ &= (E_1^T\bar y + \xi c_1,\ E_2^T\bar y + \xi c_2,\ E_1\bar x_1 + E_2\bar x_2 - \xi b,\ \bar x_1)^T.\end{aligned}$$

(i) Two multiplications of $E$: $\bar y$ on the left and for $\bar x_1, \bar x_2$ on the right,

(ii) one vector addition with $c = (c_1, c_2)$ and $b$.

Therefore, computing $A^TAv$ has four vector multiplications using $E$. In practice, a sparse matrix $E$ from the model can speed up computations.

Another key operation would be $A^TAv$. Let $\bar v = Av$, then

$$\begin{aligned}A^TAv &= \begin{bmatrix} E_1^T(E_1x_1+E_2x_2) + (c_1^Tx_1 + c_2^Tx_2 - b^Ty)c_1 \\ E_2^T(E_1x_1+E_2x_2) + (c_1^Tx_1 + c_2^Tx_2 - b^Ty)c_2 \\ E_1(E_1^Ty + z) + E_2E_2^Ty - (c_1^Tx_1 + c_2^Tx_2 - b^Ty)b \\ E_1^Ty + z \end{bmatrix}\\ &= \begin{bmatrix} (E_1^TE_1 + c_1^T)x_1 + (E_1^TE_2 + c_2^T)x_2 - (c_1b^T)y \\ (E_2^TE_1 + c_1^T)x_1 + (E_2^TE_2 + c_2^T)x_2 - (c_2b^T)y \\ -(bc_1^T)x_1 - (bc_2^T)x_2 + (E_2E_2^T + E_1E_1^T + bb^T)y + E_1z \\ E_1^Ty + z \end{bmatrix}\\ &= \begin{bmatrix} E_1^TE_1 + c_1^T & E_1^TE_2 + c_2^T & -c_1b^T & \\ E_2^TE_1 + c_1^T & E_2^TE_2 + c_2^T & -c_2b^T & \\ -bc_1^T & -bc_2^T & E_2E_2^T + E_1E_1^T + bb^T & E_1 \\ & & E_1^T & I_{n_1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ y \\ z \end{bmatrix}.\end{aligned}$$

In practice, implicitly representing the process of $A^TAv$ is better in computing software. Here we write it out to view, for theoretical interests.

### 3.1.2 Investigating the matrix

### 3.1.3 Speedy evaluations

Let $f(v) = (1/2)\|Av - d\|^2$ to be the objective function of optimization problem (3.1.2). The gradient of $f$ at $v$ is: $\nabla f(v) = A^TAv - A^Td$. Once the gradient at a point is known, the objective value at $v$, and Bregman Divergence at $u, v$ can be expressed in its gradient at $u, v$ with minimum computation overhead:

$$\begin{aligned} f(v) &= \frac{1}{2}\langle v, \nabla f(v) - A^Td\rangle + \frac{1}{2}\|d\|^2, \\ D_f(u, v) &= (1/2)\langle u - v, A^TA(u - v)\rangle \\ &= (1/2)\langle \nabla f(u) - \nabla f(v), u - v\rangle. \end{aligned}$$

This makes evaluating $\nabla f(v), f(v)$ together just as fast as evaluating $\nabla f(v)$ alone. This fact is favorable for implementations in practice. Furthermore, the difference of the function value between 2 points $v, u$ admits an interesting relation via the Bregman Divergence. Observe that $\forall u, v \in \mathbb{R}^n$ it has:

$$\begin{aligned} f(u) - f(v) &= \langle \nabla f(v), u - v\rangle + D_f(u, v) \\ &= \langle \nabla f(v), u - v\rangle + (1/2)\langle \nabla f(u) - \nabla f(v), u - v\rangle \\ &= (1/2)\langle \nabla f(u) + \nabla f(v), u - v\rangle. \end{aligned}$$

For this reason, the computation overhead for $f(u) - f(v), D_f(u, v)$ is very little as well if, $\nabla f(u), \nabla f(v)$ is already known.

## 3.2 Hoffman error bound of linear feasibility problem

## 3.3 Quadratic growth of linear feasibility problem

# Chapter 4

# Advanced Enhancement Techniques in Accelerated Proximal Gradient

This chapter is the draft for an upcoming paper. The writing style will change to fit this purpose better. It will be very terse.

**Our contributions**. We aggregate stated of the art enhancement techniques for accelerated proximal gradient method in the literatures under a unified perspective. We propose the application of conic linear feasibility problem for our algorithm and show that the convergence rate is still optimal. We conduct numerical experiment to demonstrate the relevancy of Hoffman Error bound in the practical settings for our algorithm, which is crucial for all first order methods for linear program. Using the theories of Hoffman error bound, we also demonstrate that our approach of using accelerated proximal gradient method for linear programming also yields linear convergence rate of the distance of iterates to the solution set of the best approximation problem on the linear programming KKT conditions.

There are several notable enhancements of the FISTA for function that are not necessarily strongly convex. Monotone variants are proposed by Beck [3] and Nesterov [10, 2.2.32]. Chambolle proposed the Backtracking strategy [5]. Restart is a technique can be found in Necoara et al. [9], [1] and Aujol et al. [2].

## 4.1 Preliminaries

Firstly, recall the definition of Bregman divergence $D_f(x, y)$ from Definition 1.0.1 for a differentiable function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$.

### 4.1.1 smooth plus nonsmooth weakly convex

{def:wcnvx-fxn} **Definition 4.1.1 (weakly convex function)**
*Let $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ be an l.s.c proper function. We define $F$ to be $q$ weakly convex if there exists $q \ge 0$ such that the function $F + q/2\|\cdot\|^2$ is a convex function and $q$ is the infimum of all such possible parameters.*

**Remark 4.1.2** If $q = 0$, $F$ is convex. If $F$ is weakly convex, then $F + q/2\|\cdot\|^2$ is convex and, it has $\operatorname{dom} F$ convex, and locally Lipschitz continuous on $\operatorname{ri}\operatorname{dom} F$.

{ass:sum-of-wcnvx} **Assumption 4.1.3 (sum of weakly convex smooth and nonsmooth)**
Let $F : \mathbb{R}^n \to \overline{\mathbb{R}} := f + g$ such that $f, g$ satisfy

(i) $f$ is $L$ Lipschitz smooth and $q_f$ weakly convex.

(ii) $g$ is $q_g$ weakly convex.

**Remark 4.1.4** If a function is $L$ smooth, it's $L$ weakly convex also. Here we defined $q_f$ because the actual weakly convex constant may be much smaller than $L$, and it is true in the case when $g$ is in fact convex.

{def:gm-for-ch2} **Definition 4.1.5 (gradient mapping)** *Suppose $F = f + g$ satisfies Assumption 4.1.3, define the gradient mapping for all $x \in \mathbb{R}^n$*

$$\mathcal{G}_{\beta^{-1},f,g}(x) = \beta(x - T_{\beta^{-1},f,g}(x)).$$

*If $f, g$ are clear in the context then we omit subscript and present $\mathcal{G}_\beta$.*

{lemma:mono-wcnvx-descent} **Lemma 4.1.6 (weakly convex monotone descent)**
*Let $F = f + g$ satisfies Assumption 4.1.3. Let $\bar{x} = T_{\beta^{-1},f,g}(x)$. Then, for all $x \in \mathbb{R}^n$, it has the following inequality:*

$$0 \le F(x) - F(\bar{x}) - (\beta - q_g/2 - L/2)\|x - \bar{x}\|^2.$$

*And descent is possible when $\beta \ge (L + q_g)/2$ and, it yields the descent lemma:*

$$F(\bar{x}) - F(x) \le -1/\beta\|\mathcal{G}_{1/\beta}(x)\|^2.$$

*Proof.* Use Theorem 1.0.4, set $z = x$, after some algebra it yields:

$$0 \le F(x) - F(\bar{x}) - \left(\beta - \frac{q_g + L}{2}\right)\|x - \bar{x}\|^2.$$

Using the definition of gradient mapping previously, it has for all $\beta > 0$:

$$
\begin{aligned}
0 &\le F(x) - F(\bar{x}) - \left(\beta - \frac{q_g + L}{2}\right)\|\beta^{-1}\mathcal{G}_{1/\beta}(x)\|^2 \\
&\le F(x) - F(\bar{x}) - \left(\beta^{-1} - \frac{q_g + L}{2\beta^2}\right)\|\mathcal{G}_{1/\beta}(x)\|^2
\end{aligned}
$$

Consider any $\beta \ge (q_g + L)$:

$$
\begin{aligned}
0 &\le F(x) - F(\bar{x}) - \left(\beta^{-1} - \frac{q_g + L}{2\beta^2}\right)\|\mathcal{G}_{1/\beta}(x)\|^2 \\
&\le F(x) - F(\bar{x}) + \left(\beta^{-1}/2 - \beta^{-1}\right)\|\mathcal{G}_{1/\beta}(x)\|^2 \\
&= F(x) - F(\bar{x}) - \frac{1}{2\beta}\|\mathcal{G}_{1/\beta}(x)\|^2.
\end{aligned}
$$

■

### 4.1.2 smooth plus nonsmooth convex

{ass:standard-fista}

**Assumption 4.1.7 (convex smooth and nonsmooth)** Let $F = f + g$ where $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is $L$ Lipschitz smooth, $g$ is convex and, $\operatorname*{argmin}_{x\in\mathbb{R}^n} F(x) \neq \emptyset$.

{lemma:fitsa-pg-ineq}

**Lemma 4.1.8 (proximal gradient inequality)** *If $F = f + g$ satisfies Assumption 4.1.7, then for all $x \in \mathbb{R}^n, z \in \mathbb{R}^n$, define $\bar{x} = T_{L^{-1},f,g}(x)$ it has*

$$
0 \le F(z) - F(\bar{x}) + \frac{L}{2}\|z - x\|^2 - \frac{L}{2}\|z - \bar{x}\|^2.
$$

*Proof.* Use Theorem 1.1.1 with $\mu_f = \mu_g = 0$. ■

## 4.2 FISTA made simple

This section gives convergence results under a unified perspective of accelerated proximal gradient methods with line search, backtracking, and monotone enhancements. Definition 4.2.1 unifies several combined heuristics. Theorem 4.2.5 provides a generic convergence rate for all momentum sequence satisfying Definition 4.2.3. A specialized sequence is stated in Lemma 4.2.9 which attains the lowest upper bound on the convergence rate. Theorem 4.2.11 proves the $\mathcal{O}(1/k^2)$ for optimality gap and, the norm of gradient mapping on the last iterate.

"Abstract Monotone Accelerated Proximal Gradient with line search" is "AMAPG".

{def:xxapg}

**Definition 4.2.1 (AMAPG)**
*Initialize any* $x_0, v_0 \in \mathbb{R}^n$. *Let* $(\alpha_k)_{k\ge 0}$ *be a sequence such that* $\alpha_k \in (0,1)$ $\forall k \ge 0$ *and* $\alpha_0 \in (0,1]$.

> *The algorithm makes sequences* $(x_k, v_k, y_k)_{k\ge 1}$, *such that for all* $k = 1, 2, \ldots$ *they satisfy:*
>
> $$\begin{aligned}&y_k = \alpha_k v_{k-1} + (1-\alpha_k)x_{k-1},\\ &\tilde{x}_k = T_{L_k^{-1}}(y_k),\\ &v_k = x_{k-1} + \alpha_k^{-1}(\tilde{x}_k - x_{k-1}),\\ &D_f(\tilde{x}_k, y_k) \le \frac{L_k}{2}\|\tilde{x}_k - y_k\|^2,\\ &\textit{Choose any } x_k : F(x_k) \le F(\tilde{x}_k).\end{aligned}$$

**Remark 4.2.2** Having $F(x_k) \le F(\tilde{x}_k)$ doesn't necessarily mean $F(x_k) \le F(x_{k-1})$.

The following definition characterizes the sequences $(\alpha_k)_{k\ge 0}, (\rho_k)_{k\ge 0}$, $(L_k)_{k\ge 0}$ and defines $(\beta_k)_{k\ge 0}$ for the proofs for the convergence rate.

{def:alpha-beta-rho-seq}

**Definition 4.2.3 (alpha momentum sequence)** *Let* $(\alpha_k)_{k\ge 0}$ *be a sequence in* $\mathbb{R}$ *such that* $\alpha_k \in (0,1)$ *for all* $k \in \mathbb{N}$. *Let* $(L_k)_{k\ge 0}$ *satisfy* $L_k > 0$ *for all* $k \in \mathbb{N} \cup \{0\}$. *Define the sequence* $(\rho_k)_{k\ge 0}$ *by:*

$$\rho_k = (1-\alpha_{k+1})^{-1} a_{k+1}^2 \alpha_k^{-2}.$$

*Define the sequence* $(\beta_k)_{k\ge 0}$, *let* $\beta_0 = 1$ *and, for all* $k \ge 1$ *it's defined by:*

$$\beta_k := \prod_{i=0}^{k-1}(1-\alpha_{i+1}) \max\left(1, \rho_i L_{i+1} L_i^{-1}\right).$$

{lemma:apg-iterates}

**Lemma 4.2.4 (accerated proximal gradient iterates relation)**
*The iterates* $(x_k, v_k, y_k)_{k\ge 1}$ *generated by Definition 4.2.1. Let* $z_k = \alpha_k x^+ + (1-\alpha_k)x_{k-1}$. *Then it has for all* $k \ge 1$ *that:*

$$\begin{aligned}z_k - \tilde{x}_k &= \alpha_k(x^+ - v_k)\\ x_k - y_k &= \alpha_k(x^+ - v_{k-1}).\end{aligned}$$

*Proof.* It's direct from the algorithm.

$$
\begin{aligned}
z_k - \tilde{x}_k &= (\alpha_k x^+ + (1-\alpha_k)x_{k-1}) - \tilde{x}_k \\
&= \alpha_k(x^+ + \alpha_k^{-1}(1-\alpha_k)x_{k-1} - \alpha_k^{-1}\tilde{x}_k) \\
&= \alpha_k(x^+ + \alpha_k^{-1}x_{k-1} - x_{k-1} - \alpha_k^{-1}\tilde{x}_k) \\
&= \alpha_k(x^+ + \alpha_k^{-1}(x_{k-1} - \tilde{x}_k) - x_{k-1}) \\
&= \alpha_k(x^+ - v_k), \\
z_k - y_k &= (\alpha_k x^+ + (1-\alpha_k)x_{k-1}) - (\alpha_k v_{k-1} + (1-\alpha_k)x_{k-1}) \\
&= \alpha_k(x^+ + \alpha_k^{-1}(1-\alpha_k)x_{k-1} - v_{k-1} - \alpha_k^{-1}(1-\alpha_k)x_{k-1}) \\
&= \alpha_k(x^+ - v_{k-1}).
\end{aligned}
$$

■

{thm:xxapg-fxn-cnvg} **Theorem 4.2.5 (generic AMAPG convergence)**
*Let $F = f + g$ satisfy Assumptions 4.1.7. Take the sequence $(\alpha_k)_{k\geq 0}, (\beta_k)_{k\geq 0}$ and $(\rho_k)_{k\geq 0}$ from Definition 4.2.3. Then, for all $x^+ \in \mathbb{R}^n, k \geq 1$, the convergence rate of AMAPG (Definition 4.2.1) is given by:*

$$
F(x_k) - F(x^+) + \frac{L_k\alpha_k}{2}\|x^+ - v_k\|^2 \leq \beta_k\left(F(x_0) - F(x^+) + \frac{L_0\alpha_0}{2}\|x^+ - v_0\|^2\right).
$$

*If in addition, the algorithm is initialized using line search so that $D_f(x_0, x_{-1}) \leq L_0/2\|x_0 - x_{-1}\|^2$, $\alpha_0 = 1, x_0 = v_0 = T_{L_0}x_{-1} \in \operatorname{dom} F$ and, $x^+$ is a minimizer of $F$. Then, the convergence rate simplifies:*

$$
F(x_k) - F(x^+) + \frac{L_k\alpha_k}{2}\|x^+ - v_k\|^2 \leq \frac{\beta_k L_0}{2}\|x^+ - x_{-1}\|^2.
$$

*Proof.* Define $z_k = \alpha_k x^+ + (1-\alpha_k)x_{k-1}$ for all $k \geq 1$. In the proof that follows, we list some facts in advance before their proofs which come later.

(a) Lemma 4.2.4.

(b) The sequence $(\alpha_k)_{k\geq 1}$ has for all $k \geq 1$, $\alpha_{k-1}^2\rho_{k-1}(1-\alpha_k) = \alpha_k^2$, $\alpha_k \in (0,1)$ from Definition 4.2.3.

(c) $F$ is convex and hence $F(z_k) \leq \alpha_k F(x^+) + (1-\alpha_k)F(x_{k-1})$ from Assumption 4.1.7.

(c) $F(x_k) \leq F(\tilde{x}_k)$ which is true by definition of AMAPG (Definition 4.2.1).

Now, using Theorem 1.1.1, it has for all $k \in \mathbb{N}$:

$$
0 \leq F(z_k) - F(\tilde{x}_k) - \frac{L_k}{2}\|z_k - \tilde{x}_k\|^2 + \frac{L_k}{2}\|z_k - y_k\|^2
$$

$$
\begin{aligned}
&\underset{(a)}{=} F(\alpha_k x^+ + (1-\alpha_k)x_{k-1}) - F(\tilde{x}_k) - \frac{L_k\alpha_k^2}{2}\|x^+ - v_k\|^2 + \frac{L_k\alpha_k^2}{2}\|x^+ - v_{k-1}\|^2 \\
&\underset{(c)}{\leq} \alpha_k F(x^+) + (1-\alpha_k)F(x_{k-1}) - F(\tilde{x}_k) - \frac{L_k\alpha_k^2}{2}\|x^+ - v_k\|^2 + \frac{L_k\alpha_k^2}{2}\|x^+ - v_{k-1}\|^2 \\
&= (\alpha_k - 1)F(x^+) + (1-\alpha_k)F(x_{k-1}) + F(x^+) - F(\tilde{x}_k) - \frac{L_k\alpha_k^2}{2}\|x^+ - v_k\|^2 + \frac{L_k\alpha_k^2}{2}\|x^+ - v_{k-1}\|^2 \\
&= (1-\alpha_k)(F(x_{k-1}) - F(x^+)) + \frac{L_k\alpha_k^2}{2}\|x^+ - v_{k-1}\|^2 - \left(F(\tilde{x}_k) - F(x^+) + \frac{L_k\alpha_k^2}{2}\|x^+ - v_k\|^2\right) \\
&\underset{(d)}{\leq} (1-\alpha_k)(F(x_{k-1}) - F(x^+)) + \frac{L_k\alpha_k^2}{2}\|x^+ - v_{k-1}\|^2 - \left(F(x_k) - F(x^+) + \frac{L_k\alpha_k^2}{2}\|x^+ - v_k\|^2\right) \\
&\underset{(b)}{=} (1-\alpha_k)(F(x_{k-1}) - F(x^+)) + \left(\frac{\alpha_k^2}{\alpha_{k-1}^2\rho_{k-1}}\right)\frac{L_{k-1}\alpha_{k-1}^2(\rho_{k-1}L_kL_{k-1}^{-1})}{2}\|x^+ - v_{k-1}\|^2 \\
&\quad - \left(F(x_k) - F(x^+) + \frac{L_k\alpha_k^2}{2}\|x^+ - v_k\|^2\right) \\
&= (1-\alpha_k)\left(F(x_{k-1}) - F(x^+) + \frac{L_{k-1}\alpha_{k-1}^2(\rho_{k-1}L_kL_{k-1}^{-1})}{2}\|x^+ - v_{k-1}\|^2\right) \\
&\quad - \left(F(x_k) - F(x^+) + \frac{L_k\alpha_k^2}{2}\|x^+ - v_k\|^2\right) \\
&\leq (1-\alpha_k)\left(F(x_{k-1}) - F(x^+) + \frac{L_{k-1}\alpha_{k-1}^2 \max(1, \rho_{k-1}L_kL_{k-1}^{-1})}{2}\|x^+ - v_{k-1}\|^2\right) \\
&\quad - \left(F(x_k) - F(x^+) + \frac{L_k\alpha_k^2}{2}\|x^+ - v_k\|^2\right) \\
&\leq (1-\alpha_k)\max(1, \rho_{k-1}L_kL_{k-1}^{-1})\left(F(x_{k-1}) - F(x^+) + \frac{L_{k-1}\alpha_{k-1}^2}{2}\|x^+ - v_{k-1}\|^2\right) \\
&\quad - \left(F(x_k) - F(x^+) + \frac{L_k\alpha_k^2}{2}\|x^+ - v_k\|^2\right).
\end{aligned}
$$

Unroll recursively for $k, k-1, \ldots, 0$, it implies:

$$
\begin{aligned}
0 \leq &\left(\prod_{i=0}^{k-1}(1-\alpha_{i+1})\max(1, \rho_i L_{i+1}L_i^{-1})\right)\left(F(x_0) - F(x^+) + \frac{L_0\alpha_0}{2}\|x^+ - v_0\|^2\right) \\
&- \left(F(x_k) - F(x^+) + \frac{L_k\alpha_k^2}{2}\|x^+ - v_k\|^2\right).
\end{aligned}
$$

If in addition, we assume that $x^+$ is a minimizer of $F$, and $\alpha_0 = 1, x_0 = v_0 = T_{L_0}x_{-1}$. Using

Theorem 1.1.1 it gives:

$$
\begin{aligned}
0 &\le F(x^+) - F(T_{L_{-1}} x_{-1}) - \frac{L_0}{2}\|x^+ - T_{L_0} x_{-1}\|^2 + \frac{L_0}{2}\|x^+ - x_{-1}\|^2 \\
&= F(x^+) - F(x_0) - \frac{L_0}{2}\|x^+ - v_0\|^2 + \frac{L_0}{2}\|x^+ - x_{-1}\|^2.
\end{aligned}
$$

Substituting it back to the previous inequality it yields the desired results. ■

**Remark 4.2.6** The sequence has explicit update formula:

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

{thm:xxapg-gm-cnvg} **Theorem 4.2.7 (generic AMAPG gradient mapping convergence)** *Suppose that $F = f + g$ satisfies Assumption 4.1.7. Let the sequences $(x_k, y_k, v_k)$ satisfy AMAPG (Definition 4.2.1), and take the momentum sequences $(\alpha_k)_{k\ge 0}, (\beta_k)_{k\ge 0}, (\rho_k)_{k\ge 0}$ from Definition 4.2.3. If in addition,*

(i) *The sequence $(\alpha_k)_{k\ge 0}$ has $\alpha_0 = 1$ and, AMAPG is initialized with $L_0 \ge L$ or, equivalently a successful line search satisfying $D_f(x_0, x_{-1}) \le L_0/2\|x_0 - x_{-1}\|^2$;*

(ii) *$v_0 = x_0 = T_{1/L_0, f, g}(x_{-1})$ for any $x_{-1} \in \mathbb{R}^n$ and there exists $x^+$ which is a minimizer of $F$.*

*Then, we have the convergence of gradient mapping, it satisfies for all $k \ge 1$ the inequality:*

{ineq:xxapg-gm-cnvg-prt1}
$$
\|\mathcal{G}_{1/L_k}(y_k)\| \le \sqrt{\beta_k} L_k L_0 \left(1 - \min(\rho_{k-1}, L_k^{-1} L_{k-1})^{1/2}\right)\|x^+ - v_0\|. \tag{4.2.1}
$$

*It has also:*

{ineq:xxapg-gm-cnvg-prt2}
$$
\frac{1}{2L_0}\|\mathcal{G}_{1/L_0}(x_{-1})\|^2 \le F(x_{-1}) - F(x_0). \tag{4.2.2}
$$

*Proof.* (4.2.2) is direct because $x_0 = T_{1/L_0, f, g}(x_{-1})$ and $D_f(x_0, x_{-1}) \le L_0/2\|x_0 - x_{-1}\|^2$ is assumed in the statement hypothesis, using Lemma 4.1.8 with $x = x_{-1}, z = x_{-1}$, by Definition 4.1.5 it has

$$
\begin{aligned}
0 &\le F(x_{-1}) - F(x_0) + 0 - \frac{L_0}{2}\|x_{-1} - x_0\|^2 \\
&= F(x_{-1}) - F(x_0) - \frac{L_0}{2}\|L_0^{-1}\mathcal{G}_{1/L_0}(x_{-1})\|^2.
\end{aligned}
$$

We label the following results prior to their proofs which will come later for a sleeker exposition for the proof of (4.4.2).

(a) From Definition 4.2.1, the gradient mapping satisfies for all $k \ge 1$ that $\|\mathcal{G}_{1/L_k}(y_k)\| = L_k\alpha_k\|v_k - v_{k-1}\|$.

(b) We have $(a_k)_{k\ge 1}$ satisfying $\forall k \ge 1$ that $(1-\alpha_k)\rho_{k-1} = \alpha_k^2/\alpha_{k-1}^2$ from the statement hypothesis. We assumed $\alpha_0 = 0, \beta_0 = 1$, $x^+$ is a minimizer of $F$ and, a successful line search in item (i). Then using these it has for all $k \ge 0$ it has $\frac{\alpha_k}{\sqrt{\beta_k L_0}}\|x^+ - v_k\| \le \|x^+ - v_0\|$.

(c) The sequence $(\alpha_k)_{k\ge 0}$ has $(1-\alpha_k)\rho_{k-1} = \alpha_k^2/\alpha_{k-1}^2$ from the statement hypothesis so $\alpha_k/\alpha_{k-1} = \sqrt{\rho_{k-1}(1-\alpha_k)}$ for all $k \ge 1$.

(d) The definition of $(\beta_k)_{k\ge 0}$ from Definition 4.2.3.

Using the above intermediate results, the convergence in (4.2.1) can be derived. From (a) it has for all $k \ge 0$:

$$
\begin{aligned}
\|\mathcal{G}_{1/L_k}(y_k)\| &= L_k\alpha_k\|v_k - v_{k-1}\| \\
&\le L_k\alpha_k(\|v_k - x^+\| + \|v_{k-1} - x^+\|) \\
&\underset{\text{(b)}}{\le} L_k\alpha_k\left(\frac{\sqrt{\beta_k L_0}}{\alpha_k}\|x^+ - v_0\| + \frac{\sqrt{\beta_{k-1}L_0}}{\alpha_{k-1}}\|x^+ - v_0\|\right) \\
&= L_k\sqrt{L_0}\left(\sqrt{\beta_k} + \frac{\alpha_k\sqrt{\beta_{k-1}}}{\alpha_{k-1}}\right)\|x^+ - v_0\| \\
&= \sqrt{\beta_k L_0}L_k\left(1 + \frac{\alpha_k}{\alpha_{k-1}}\sqrt{\frac{\beta_{k-1}}{\beta_k}}\right)\|x^+ - v_0\| \\
&\underset{\text{(d)}}{=} \sqrt{\beta_k L_0}L_k\left(1 + \frac{\alpha_k}{\alpha_{k-1}}\left((1-\alpha_k)\max(1, \rho_{k-1}L_kL_{k-1}^{-1})\right)^{-1/2}\right)\|x^+ - v_0\| \\
&\underset{\text{(c)}}{=} \sqrt{\beta_k L_0}L_k\left(1 + ((1-\alpha_k)\rho_{k-1})^{1/2}\left((1-\alpha_k)\max(1, \rho_{k-1}L_kL_{k-1}^{-1})\right)^{-1/2}\right)\|x^+ - v_0\| \\
&= \sqrt{\beta_k L_0}L_k\left(1 + \left(\rho_{k-1}^{-1}\max(1, \rho_{k-1}L_kL_{k-1}^{-1})\right)^{-1/2}\right)\|x^+ - v_0\| \\
&= \sqrt{\beta_k L_0}L_k\left(1 + \max(\rho_{k-1}^{-1}, L_kL_{k-1}^{-1})^{-1/2}\right)\|x^+ - v_0\| \\
&= \sqrt{\beta_k L_0}L_k\left(1 + \min(\rho_{k-1}, L_k^{-1}L_{k-1})^{1/2}\right)\|x^+ - v_0\|.
\end{aligned}
$$

Now, **let's proof intermediate results (a)**. From the Definition 4.2.1 it has

$$
y_k = \alpha_k v_{k-1} + (1-\alpha_k)x_{k-1} \iff v_{k-1} = \alpha_k^{-1}(y_k - (1-\alpha_k)x_{k-1}).
$$

Current iterates $v_k$ is updated via $x_{k-1}, \tilde{x}_k$ so consider:

$$
\begin{aligned}
v_k - v_{k-1} &= (x_{k-1} + \alpha_k^{-1}(\tilde{x}_k - x_{k-1})) - \alpha_k^{-1}(y_k - (1-\alpha_k)x_{k-1}) \\
&= x_{k-1} + \alpha_k^{-1}(\tilde{x}_k - x_{k-1}) - \alpha_k^{-1}y_k + (\alpha_k^{-1} - 1)x_{k-1} \\
&= \alpha_k^{-1}(\tilde{x}_k - x_{k-1}) - \alpha_k^{-1}y_k + \alpha_k^{-1}x_{k-1} \\
&= \alpha_k^{-1}\tilde{x}_k - \alpha_k^{-1}y_k = \alpha_k^{-1}(\tilde{x}_k - y_k) = \alpha_k^{-1}(T_{1/L_k}y_k - y_k) \\
&= -\alpha_k^{-1}L_k^{-1}(\mathcal{G}_{1/L_k}(y_k)).
\end{aligned}
$$

**We now prove result (b)**. The base case $k = 1$ is verified by the assumption that $x_0 = v_0 = T_{L_0}x_{-1}$. Apply Lemma 4.1.8 with $z = x^+$ as a minimizer it yields:

$$
\begin{aligned}
& 0 \le F(x^+) - F(T_{L_{-1}}x_{-1}) - \frac{L_0}{2}\|x^+ - T_{L_0}x_{-1}\|^2 + \frac{L_0}{2}\|x^+ - x_{-1}\|^2 \\
&= F(x^+) - F(x_0) - \frac{L_0}{2}\|x^+ - v_0\|^2 + \frac{L_0}{2}\|x^+ - x_{-1}\|^2 \\
&\le -\frac{L_0}{2}\|x^+ - v_0\|^2 + \frac{L_0}{2}\|x^+ - x_{-1}\|^2 \\
\implies & 0 \le \frac{L_0}{2}\left(\|x^+ - x_{-1}\| - \|x^+ - v_0\|\right).
\end{aligned}
$$

Because $\beta_0 = \alpha_0 = 1$, the base case holds. For all $k \ge 1$, we consider the convergence claim and use the assumption that $x^+$ is a minimizer of $F$ so, it has from Theorem 4.2.5 that

$$
\begin{aligned}
& 0 \le \frac{L_0\beta_k}{2}\|x^+ - x_{-1}\|^2 - F(x_k) + F(x^+) - \frac{L_k\alpha_k^2}{2}\|x^+ - v_k\|^2 \\
&\le \frac{L_0\beta_k}{2}\|x^+ - x_{-1}\|^2 - \frac{L_k\alpha_k^2}{2}\|x^+ - v_k\|^2 \\
&= \frac{\alpha_k^2 L_k}{2}\left(\frac{\beta_k}{\alpha_k^2 L_0}\|x^+ - x_{-1}\|^2 - \|x^+ - v_k\|^2\right) \\
\iff & 0 \le \|x^+ - x_{-1}\| - \frac{\alpha_k}{\sqrt{\beta_k L_0}}\|x^+ - v_k\|.
\end{aligned}
$$

■

**Remark 4.2.8** The above theorem is improved from Alamo et al. [1].

{lemma:xxapg-seq-bnd} **Lemma 4.2.9 (specialized AMAPG momentum sequence)**
*Take sequences $(\alpha_k)_{k\ge0}, (\rho_k)_{k\ge0}, (\beta_k)_{k\ge0}$ as in Definition 4.2.3. In addition, assume that $\alpha_0 = 1$. If, we set $\rho_{k-1} = L_k^{-1}L_{k-1}$ such that $L_k > 0$ for all $k \ge 1$, then for all $k \ge 1$, the sequence $(\beta_k)_{k\ge0}$ has the inequality:*

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

*Proof.* We state the following intermediate results needed to construct the proof. They will be proved at the end.

(a) $(\beta_k)_{k\geq 0}$ is monotone decreasing, and it's strictly larger than zero.

(b) Because $\rho_k L_{k+1} L_k^{-1} = 1$ for all $k \geq 0$, the definition of $(\beta_k)_{k\geq 0}$ simplifies and $\beta_k = (\alpha_k^2/\alpha_0^2)(L_k/L_0)$. As a consequence it also gives for all $k \geq 1$ that:

$$
\begin{aligned}
\alpha_k^2 &= \alpha_0^2 \beta_k L_0 L_k^{-1}, \\
\alpha_k &= 1 - \beta_k/\beta_{k-1}.
\end{aligned}
$$

Starting with results (b), and combine the two equality it gives for all $k \geq 1$ the equality

$$
\begin{aligned}
& 0 = (1 - \beta_k/\beta_{k-1})^2 - \alpha_0^2 L_0 L_k^{-1} \beta_k \\
\iff & 0 = (1 - \beta_k/\beta_{k-1}) - \alpha_0 \sqrt{L_0 L_k^{-1} \beta_k} \\
\iff & 0 = (\beta_k^{-1} - \beta_{k-1}^{-1}) - \alpha_0 \sqrt{L_0 L_k^{-1} \beta_k^{-1}} \\
& = \left(\sqrt{\beta_k^{-1}} + \sqrt{\beta_{k-1}^{-1}}\right)\left(\sqrt{\beta_k^{-1}} - \sqrt{\beta_{k-1}^{-1}}\right) - \alpha_0 \sqrt{L_0 L_k^{-1} \beta_k^{-1}} \\
& \underset{\text{(a)}}{\leq} 2\sqrt{\beta_k^{-1}}\left(\sqrt{\beta_k^{-1}} - \sqrt{\beta_{k-1}^{-1}}\right) - \alpha_0 \sqrt{L_0 L_k^{-1} \beta_k^{-1}} \\
\iff & 0 \leq 2\left(\sqrt{\beta_k^{-1}} - \sqrt{\beta_{k-1}^{-1}}\right) - \alpha_0 \sqrt{L_0 L_k^{-1}}.
\end{aligned}
$$

Since this is true for all $k \geq 1$, taking a telescoping sum of the above series gives

$$
\begin{aligned}
0 &\leq \left(\sum_{i=1}^{k} \sqrt{\beta_i^{-1}} - \sqrt{\beta_{i-1}^{-1}}\right) - \sum_{i=1}^{k} \frac{\alpha_0}{2}\sqrt{L_0 L_k^{-1}} \\
&= \sqrt{\beta_k^{-1}} - \sqrt{\beta_0^{-1}} - \frac{\alpha_0\sqrt{L_0}}{2}\sum_{i=1}^{k}\sqrt{L_k^{-1}} \\
&= \sqrt{\beta_k^{-1}} - 1 - \frac{\alpha_0\sqrt{L_0}}{2}\sum_{i=1}^{k}\sqrt{L_k^{-1}}.
\end{aligned}
$$

Therefore, transforming the inequality it has:

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

**Let's now justify (a)**. When $\rho_i = L_{i+1}L_i^{-1}$, the big product simplifies and, it has:

$$
\beta_k = \prod_{i=0}^{k-1}(1 - \alpha_{i+1}) = (1 - \alpha_k)\beta_{k-1}.
$$

Since $\alpha_k \in (0,1)$, $\beta_k$ is monotonically decreasing. **To see (b)**, it has from the above which also justifies $1 - \alpha_k = \beta_k/\beta_{k-1}$. Recall that sequence $(\alpha_k)_{k\ge 0}$ has $\forall k \ge 1$ that $\alpha_{k-1}^2\rho_{k-1}(1-\alpha_k) = \alpha_k^2$, using it we can simplify the product for $\beta_k$, it follows that

$$\begin{aligned}\beta_k &= \prod_{i=0}^{k-1}(1-\alpha_{i+1}) = \prod_{i=1}^{k}\alpha_i^2\alpha_{i-1}^{-2}\rho_{i-1}^{-1} = \prod_{i=1}^{k}\alpha_i^2\alpha_{i-1}^{-2}L_i^{-1}L_{i-1} \\ &= \left(\frac{\alpha_k^2\alpha_{k-1}^2\ldots\alpha_1^2}{\alpha_{k-1}^2\alpha_{k-2}^2\ldots\alpha_0^2}\right)\left(\frac{L_kL_{k-1}\ldots L_1}{L_{k-1}L_{k-1}\ldots L_0}\right) = \frac{\alpha_k^2}{\alpha_0^2}\frac{L_k}{L_0}.\end{aligned}$$

Rearranging it gives: $\alpha_0^2L_0\beta_kL_k^{-1} = \alpha_k^2$. ■

**Remark 4.2.10** The technique of the proof we used here is very similar to Güler [6, Lemma 2.2]. A simpler upper bound is more practical. For all $k \ge 1$ let

$$\widehat{L}_k = \max\left(L_0, \left(\frac{1}{k}\sum_{i=1}^{k}\sqrt{L_i^{-1}}\right)^{-2}\right).$$

Then,

$$\begin{aligned}\beta_k &\le \left(1 + \frac{\alpha_0\sqrt{L_0}}{2}\sum_{i=1}^{k}\sqrt{L_k^{-1}}\right)^{-2} \le \left(1 + \frac{1}{2}\alpha_0\sqrt{L_0}k\sqrt{\widehat{L}_k^{-1}}\right)^{-2} \\ &= \left(1 + \frac{k\alpha_0\sqrt{L_0\widehat{L}_k^{-1}}}{2}\right)^{-2} = L_0^{-1}\widehat{L}\left(\sqrt{L_0^{-1}\widehat{L}_k} + \frac{k\alpha_0}{2}\right)^{-2} \\ &\le L_0^{-1}\widehat{L}_k\left(1 + \frac{k\alpha_0}{2}\right)^{-2} = \frac{4\widehat{L}_k}{L_0(2+k\alpha_0)^2}.\end{aligned}$$

This simplifies the convergence claim back in Theorem 4.2.5.

{thm:xxapg-specialized-cnvg}

**Theorem 4.2.11 (specialized AMAPG convergence rate)** *Suppose that $F = f + g$ satisfy Assumption 4.1.7. Let the sequences $(x_k, v_k, v_k)_{k\ge 0}$ and $(L_k)_{k\ge 0}$ satisfy AMAPG in Definition 4.2.1 and, assume that the AMAPG is initialized by $x_0 = v_0 = T_{1/L_0}(x_{-1})$ and, assume $\rho_{k-1} = L_k^{-1}L_k$, $\alpha_0 = 1$ so the sequence $(\alpha_k)_{k\ge 0}$ satisfies for all $k \ge 1$: $\alpha_{k-1}^2L_k^{-1}L_{k-1}(1-\alpha_k) = \alpha_k^2$. Let $x^+$ be a minimizer of $F$, define*

$$\widehat{L}_k := \max\left(L_0, \left(\frac{1}{k}\sum_{i=1}^{k}\sqrt{L_i^{-1}}\right)^{-2}\right).$$

*Then, we have convergence claim:*

(i)

$$F(x_k) - F(x^+) + \frac{L_k \alpha_k}{2}\|x^+ - v_k\|^2 \le \frac{2\widehat{L}_k}{(2+k)^2}\|x^+ - x_{-1}\|^2.$$

(ii)

$$\|\mathcal{G}_{1/L_k}(y_k)\| \le \frac{2\widehat{L}_k L_k}{2+k}\left(1 - L_k^{-1/2} L_{k-1}^{1/2}\right)\|x^+ - v_0\|.$$

*Proof.* To see (i), use Lemma 4.2.9 and its remark to bound $(\beta_k)_{k\ge 1}$ and then, apply Theorem 4.2.5 because the assumptions of $x^+, (\alpha_k)_{k\ge 0}, (\rho_k)_{k\ge 0}$ suit. To see (ii), the convergence claim from 4.2.7 simplifies with $\widehat{L}_k \ge L_0$ and, it has

$$\|\mathcal{G}_{1/L_k}(y_k)\| \le \left(\frac{2\sqrt{\widehat{L}_k L_0} L_k}{2+k}\right)\left(1 + \min(\rho_{k-1}, L_k^{-1} L_{k-1})^{1/2}\right)\|x^+ - v_0\| \tag{4.2.3}$$

$$= \left(\frac{2\sqrt{\widehat{L}_k L_0} L_k}{2+k}\right)\left(1 + L_k^{-1/2} L_{k-1}^{1/2}\right)\|x^+ - v_0\| \tag{4.2.4}$$

$$\le \left(\frac{2\widehat{L}_k L_k}{2+k}\right)\left(1 + L_k^{-1/2} L_{k-1}^{1/2}\right)\|x^+ - v_0\|. \tag{4.2.5}$$

■

## 4.3 Algorithmic description of AMAPG

There are several components to the AMAPG algorithm. This section will introduce various type of implementations that can be fitted into AMAPG in Definition 4.2.1.

### 4.3.1 Line search routines

**Algorithm 1** Armijo Line Search

**Function ArmijoLS:**

| | |
|---|---|
| $f : \mathbb{R}^n \to \mathbb{R}$ | Convex Lipschitz smooth |
| $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ | Convex |
| $x \in \mathbb{R}^n$ | Vector |
| $v \in \mathbb{R}^n$ | Vector |
| $L \in \mathbb{R}$ | $L > 0$ |
| $\alpha \in \mathbb{R}$ | $\alpha \in (0, 1]$ |
| $\cdots$ | Ignore extra inputs |

$\alpha^+ := (1/2)\left(\alpha\sqrt{\alpha^2 + 1} - \alpha^2\right)$.

{alg:armijo-ls-yplus} $y^+ := \alpha^+ v + (1 - \alpha^+)x$.

$L^+ := L$.

**for** $i = 1, 2, \ldots, 53$ **do**

  $L^+ := 2L^+$.

{alg:armijo-ls-xplus}  $x^+ := T_{1/L^+,f,g}(y^+)$.

  **if** $D_f(x^+, y^+) \le (L^+/2)\|x^+ - y^+\|^2$ **then**

    **break**

  **end if**

  $L^+ := 2^i L$

**end for**

{alg:armijo-ls} **Return:** $x^+, y^+, \alpha^+, L^+$

Algorithm 1 performs a step of Armijo line search and a step of accelerated proximal gradient. The function can be used for each iteration in the inner loop of the algorithm. Here are the explanations for all its input parameters:

(i) $f, g$ are functions satisfying Assumption 4.1.7.

(ii) $x, v$ are the $x_k, v_k$ iterates in Definition 4.2.1.

(iii) $\alpha$ are the current $\alpha_k$ in Definition 4.2.1.

(iv) $L$ is the estimate of the Lipschitz constant of $f$ passed in by the inner loop.

Iterates $x^+, y^+$ and parameters $\alpha^+, L^+$ are returned to the callers at the end.

**Algorithm 2** Chambolle's Backtracking

**Function ChamBT Inputs:**

| | |
|---|---|
| $f : \mathbb{R}^n \to \mathbb{R}$ | Convex Lipschitz smooth |
| $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ | Convex |
| $x \in \mathbb{R}^n$ | Vector |
| $v \in \mathbb{R}^n$ | Vector |
| $L \in \mathbb{R}$ | Number, $L > 0$ |
| $\alpha \in \mathbb{R}$ | Vector |
| $L_{\min} \in \mathbb{R}$ | Number, $L_{\min} > 0$ |
| $\rho \in \mathbb{R}$ | Number, $\rho \in (0, 1)$ |

$L^+ := \max(L_{\min}, \rho L)$.
**for** $i = 1, 2, \ldots, 53$ **do**
    $\alpha^+ := (1/2)\left(\alpha\sqrt{\alpha^2 + L/L^+} - \alpha^2\right)$.
    $y^+ := \alpha^+ v + (1 - \alpha^+)x$.
    $x^+ := T_{1/L^+, f, g}(y^+)$.
    **if** $2D_f(x^+, y^+) \le \|x^+ - y^+\|^2$ **then**
        **break**
    **end if**
    $L^+ := 2^i L^+$.
**end for**
{alg:chambolle-btls} **Return:** $x^+, \alpha^+, L^+$

Algorithm 2 attempts to decrease the Lipschitz estimate $L_k$ for $f$ in an iteration of the inner loop. The above implementations were adapted and simplified from Chambolle et al. [5]. It takes in additional parameters $L_{\min}, \rho$ compared to Algorithm 1. Here are their explanations:

(i) $L_{\min}$ determines a lower bound of Lipschitz estimates. It's the lowest value of an estimate $L_k$ allowed. It increases stability of the algorithm by preventing unnecessary triggering a line search routine to recovers from an underestimated $L_k$ that doesn't satisfy the Lipschitz smoothness condition for $f$ at the current iterate.

(ii) $\rho \in (0, 1)$ is the decay ratio. It's use to shrink the current estimate of $L$ and produce $L^+$ at the start of the forloop before verifying the smoothness condition.

### 4.3.2 Monotone routines

**Algorithm 3** Beck's monotone routine

**Function BeckMono Inputs:**

| | |
|---|---|
| $f : \mathbb{R}^n \to \mathbb{R}$ | Convex Smooth |
| $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ | Convex |
| $\tilde{x} \in \mathbb{R}^n$ | Vector |
| $x \in \mathbb{R}^n$ | Vector |
| $\rho$ | Number $\rho \in (0, 1)$ Number |
| $G \in \mathbb{R}$ | Number |

$x^+ = \text{argmin}\{(f + g)(z) : z \in \{\tilde{x}, x\}\}$.
**Return:** $x^+, \eta, G$

{alg:beck-mono}

Algorithm 3 is a subroutine for asserting monotone condition on function value. The parameter $G$ has no actual usage besides making it compatible with Algorithm 4 in the context of Algorithm 5. The input $\tilde{x}$ is the candidate iterate produced by FISTA without monotone constraints and $x$ is the previous iterates $x_{k-1}$ in the inner loop.

**Algorithm 4** Nesterov's monotone routine

**Function NesMono Inputs:**

| | |
|---|---|
| $f : \mathbb{R}^n \to \mathbb{R}$ | Lipschitz Smooth |
| $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ | Weakly Convex |
| $\tilde{x} \in \mathbb{R}^n$ | Vector |
| $x \in \mathbb{R}^n$ | Vector |
| $\eta \in \mathbb{R}$ | Number $\eta > 0$ |
| $G \in \mathbb{R}$ | Number |

$\hat{y} := \text{argmin}\{(f + g)(z), x \in \{\tilde{x}, x\}\}$.
$x^+ := T_{1/\eta}(\hat{y})$.
**for** i = 1, 2, . . . , 53 **do**
  **if** $(f + g)(x^+) - (f + g)(\hat{y}) \le -1/(2\eta)\|\mathcal{G}_{1/\eta}(\hat{y})\|^2$ **then**
    **Break**
  **end if**
  $\eta := 2\eta$.
  $x^+ := T_{1/\eta}(\hat{y})$.
**end for**
$G := \eta(x^+ - \hat{y})$.
**return:** $x^+, \eta, G$

{alg:nes-mono}

The above Algorithm 4 implements and adapts Nesterov's monotone scheme from Nesterov [10, 2.2.32] for AMAPG. In addition to Algorithm 3, $\eta$ is a new input parameter and $G$ has a significance role. $\eta$ is a stepsize parameter for weakly convex objective $F = f + g$ satisfying Assumption 4.1.3. $G$ is the norm of the gradient mapping updated at $\hat{y}$ which will be returned to the inner loop to verify exit conditions.

### 4.3.3 AMAPG main algorithm

**Algorithm 5** AMAPG main alorithm

**Function AMAPG Inputs:**

| | |
|---|---|
| $f : \mathbb{R}^n \to \mathbb{R}$ | Lipschitz Smooth |
| $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ | Weakly Convex |
| $x_{-1}$ | Vector |
| $L \in \mathbb{R}$ | $L > 0$ |
| $r$ | $r \in (0, 1)$ |
| $\rho \in \mathbb{R}$ | $\rho \in (0, 1)$ |
| $N_{\min} \in \mathbb{N}$ | $N \geq 1$ |
| $N \in \mathbb{N}$ | $N \geq N_{\min}$ |
| $\epsilon \in \mathbb{R}$ | Number |
| $\mathbf{L}$ | Algorithm 1 or 2 |
| $\mathbf{M}$ | Algorithm 3 or 4 |
| $\mathbf{E}_\chi$ | Exit Condition |

$\alpha_0 := 1.$
{alg:xxapg-armijo} $x_0, y_0, \alpha_1, L_0 :=$ **ArmijoLS**$(f, g, x_{-1}, x_{-1}, L, \alpha_0)$.
{alg:xxapg-gzero} $\eta_0 := L_0; v_0 := x_0; G_0 = \|\sqrt{L_0}(x_0 - y_0)\|$.
**if** $G_0 \leq \epsilon$ **then**
{alg:xxapg-exit1}     **Return:** $x_k, 0, L_0, G_0$
**end if**
$\overline{L} := L_0$.
**for** $k := 1, 2, \ldots, N$ **do**
    $\tilde{x}_k, y_k, \alpha_{k+1}, L_k := \mathbf{L}(f, g, v_{k-1}, x_{k-1}, L_{k-1}, \alpha_k, r\overline{L}, \rho)$.
    $v_k := x_{k-1} + \alpha_k^{-1}(\tilde{x}_k - x_{k-1})$.
{alg:xxapg-lbar}     $\overline{L} := \max(L_k, L_{k-1})$.
{alg:xxapg-rhosqrt}     $\rho := \rho^{1/2}$ **if** $L_k > L_{k-1}$ **else** $\rho$.
    $G_k := \|\sqrt{L_k}(\tilde{x}_k - y_k)\|$
    $x_k, \eta_{k+1}, G_k^+ := \mathbf{M}(f, g, \tilde{x}_k, x_{k-1}, \eta_k, G_k)$.
    **if** $G_k^+ < \epsilon$ **or** $(\mathbf{E}_\chi$ **and** $k \geq N_{\min})$ **then**
{alg:xxapg-exit2}         **break**
    **end if**
**end for**
{alg:xxapg} **Return:** $x_k, k, \overline{L}, G_k^+$

The above Algorithm 5 is an implementation of AMAPG in Definition 4.2.1. The first iterates $x_0$ is produced by a step of proximal gradient descent through Algorithm 1 so, it has $x_0 = v_0 = T_{1/L_0}(x_{-1})$, and consequently all results from Theorem 4.2.11 apply.

There are several tricks involved with it and it deserves explaination. At line 12 it keeps the largest Lipschitz estimate from the line search routines under the variable $\overline{L}$ and, the algorithm returns it after exiting the for loop, and it's used as an input for routine $\mathbf{L}$ to determine the lower bound of the Lipschitz estimated which is exclusived by Chambolle's backtracking (Algorithm 2). Whenever it detects that $L_k > L_{k-1}$, i.e: the estimated Lipschitz constant had increased, it takes the square root of the decay ratio, making it closer to one. This decay ratio parameter is exclusively utilised by Chambolle's backtracking routine (Algorithm 2). This trick prevents triggering backtracking routine frequently if $\rho$ is a small number.

Parameters $r, \rho$ are chosen in the discretion of the practitioners. For example, we chose $r = 0.4, \rho = 2^{1/1024}$.

{obs:xxapg-exit-cond}

**Observation 4.3.1 (AMAPG exit conditions)** *Algorithm 5 exits and returns its results at line 6, or at line 17. If exited, then at least one of the conditions are true.*

(i) $\|\sqrt{L_0}(x_0 - x_{-1})\| \le \epsilon$, *and the line search on* $x_{-1}$ *is successful so* $D_f(x_0, x_{-1}) \le L_0/2\|x_0 - x_{-1}\|^2$. *It has* $y_0 = x_{-1}$ *because it passes* $\alpha_0 = 1$ *at line 3.*

(ii) $G_k^+ \le \epsilon$ *or,* $\mathbf{E}_\chi$ *is true and* $k > N_{\min}$.

## 4.4 Examples of AMAPG in the literature

**Example 4.4.1 (MFISTA with Armijo line search)**

**Algorithm 6** MFISTA with Armijo Line Search

**Input:** $x_{-1} \in \mathbb{R}^n, L_0 \in \mathbb{R}^n, f : \mathbb{R}^n \to \mathbb{R}, g : \mathbb{R}^n \to \overline{\mathbb{R}}$
$x_0 := y_0, t_0 := 1.$
**for** $k = 0, 1, 2, \ldots$ **do**
    $\tilde{x}_{k+1} := T_{L_k^{-1}}(y_k).$
    **if** $D_f(\tilde{x}_{k+1}, y_k) > L_k/2\|\tilde{x}_{k+1} - y_k\|^2$ **then**
        $L_k := \operatorname*{argmin}_{i=1,2,\ldots}\left\{i : D_f(T_{2^{-i}L_k^{-1}}(y_k), y_k) \le 2^{i-1}L_k \left\|T_{2^{-i}L^{-1}}y_k - y_k\right\|^2\right\}.$
        $\tilde{x}_{k+1} := T_{L_k^{-1}}y_k.$
    **end if**
    Choose $x_{k+1} \in \{\tilde{x}_{k+1}, x_k\}$ such that $F(x_{k+1}) \le \min(F(x_k), F(\tilde{x}_{k+1}))$.
    $t_{k+1} := (1/2)\left(1 + \sqrt{1 + 4t_k^2}\right).$
    $y_{k+1} := x_{k+1} + t_k t_{k+1}^{-1}(\tilde{x}_{k+1} - x_{k+1}) + (t_k - 1)t_{k+1}^{-1}(x_{k+1} - x_k).$
{alg:mfista-armijo} **end for**

We now demonstrate that Algorithm 6 is a special case of Definition 4.2.1. Let's consider $y_{k+1}$ produced the AMAPG. If $x_k = x_{k-1}$ then replacing all instance of $x_k$ by $x_{k-1}$ it has:

$$
\begin{aligned}
y_{k+1} &= \alpha_{k+1}(v_k) + (1 - \alpha_{k+1})x_{k-1} \\
&= \alpha_{k+1}(x_{k-1} + \alpha_k^{-1}(\tilde{x}_k - x_{k-1})) + (1 - \alpha_{k+1})x_{k-1} \\
&= \alpha_{k+1}x_{k-1} + \alpha_{k+1}\alpha_k^{-1}(\tilde{x}_k - x_{k-1}) + (1 - \alpha_{k+1})x_{k-1} \\
&= x_{k-1} + \alpha_{k+1}\alpha_k^{-1}(\tilde{x}_k - x_{k-1})
\end{aligned}
$$

Similarly when $x_k = \tilde{x}_k$ it produces:

$$
\begin{aligned}
y_{k+1} &= \alpha_{k+1}v_k + (1 - \alpha_{k+1})\tilde{x}_k \\
&= \alpha_{k+1}(x_{k-1} + \alpha_k^{-1}(\tilde{x}_k - x_{k-1})) + (1 - \alpha_{k+1})x_k \\
&= \alpha_{k+1}\left((1 - \alpha_k^{-1})x_{k-1} + (\alpha_k^{-1} - 1)\tilde{x}_k + \tilde{x}_k\right) + (1 - \alpha_{k+1})\tilde{x}_k \\
&= \alpha_{k+1}\left((\alpha_k^{-1} - 1)(\tilde{x}_k - x_{k-1}) + \tilde{x}_k\right) + (1 - \alpha_{k+1})\tilde{x}_k. \\
&= \tilde{x}_k + \alpha_{k+1}(\alpha_k^{-1} - 1)(\tilde{x}_k - x_{k-1}).
\end{aligned}
$$

Let's denote $y'_k, x'_k, \tilde{x}'_k$ as the $y_k, x_k, \tilde{x}_k$ produced by Algorithm 6. Observe that if $x'_0$ is not the minimizer then it has $\tilde{x}'_1 = T_{L_0^{-1}}(y'_0) = T_{L_0^{-1}}(x'_0)$. Then $F(\tilde{x}'_1) < F(x'_0)$ is true. So $x'_1 = \tilde{x}_1 = T_{L_0^{-1}}(x'_0)$. Since $t_0 = 1$, it has $y'_1 = \tilde{x}'_1 + (t_0 - 1)t_1^{-1}(\tilde{x}'_1 - x'_0) = \tilde{x}'_1$.

Summarize the above results compactly, it has for all $k \ge 0$

{eqn:emp:result-item-1}

$$
y_{k+1} = \begin{cases}
x_{k-1} + \alpha_{k+1}\alpha_k^{-1}(\tilde{x}_k - x_{k-1}) & \text{if } x_k = x_{k-1} \wedge k \ge 1, \\
\tilde{x}_k + \alpha_{k+1}(\alpha_k^{-1} - 1)(\tilde{x}_k - x_{k-1}) & \text{if } x_k = \tilde{x}_k \wedge k \ge 1, \\
\alpha_1 v_0 + (1 - \alpha_1)x_0 & \text{if } k = 0.
\end{cases} \tag{4.4.1}
$$

Then it has for all $k \ge 0$:

{eqn:emp:result-item-2}

$$
y'_{k+1} = \begin{cases} x'_k + t_k t_{k+1}^{-1}(\tilde{x}_{k+1} - x_k) & \text{if } x'_{k+1} = x'_k \wedge k \ge 1, \\ x'_{k+1} + (t_k - 1)t_{k+1}^{-1}(\tilde{x}'_{k+1} - x'_k) & \text{if } x'_{k+1} = \tilde{x}'_{k+1} \wedge k \ge 1, \\ \tilde{x}'_1 & \text{if } k = 0. \end{cases} \tag{4.4.2}
$$

Let $x_{-1} \in \mathbb{R}^n$. If we choose $v_0 = x_0 = T_{L_0^{-1}} x_{-1}$, then $y_1 = \alpha_1 x_0 + (1-\alpha_1)x_0 = x_0 = T_{L_0^{-1}}(x_{-1})$. Next, we make $\alpha_k^{-1} = t_k$, then (4.4.1), (4.4.2) are equivalent.

## 4.5 Practical enhancement from the Nesterov's Monotone Variant

Under Assumption 4.1.3, Theorem 4.5.2 shows the Nesterov's monotone algorithm in Definition 4.5.1 eventually terminate.

{def:nes-monotone-scheme}

**Definition 4.5.1 (nonconvex Nesterov's monotone scheme)**
*Suppose $F = f + g$ satisfies Assumption 4.1.3. Let $L_0 \ge L$. Let $(\alpha_k)_{k\ge 0}$ with $\alpha_0 = 1$ and, it satisfies for all $k \ge 1$: $L_k^{-1} L_{k-1}\alpha_{k-1}^2(1-\alpha_k) = \alpha_k^2$. Initialize the algorithm with $\hat{y}_0 = v_0 = x_0 = T_{1/L_0}(x_{-1})$, $\eta_0 = L_0$, for some $x_{-1} \in \mathbb{R}^n$ and $L_0$ such that $F(x_0) \le F(x_{-1})$. The algorithm is defined by sequences $(y_k, v_k, x_k)_{k\ge 1}$ and $(\tilde{x}_k, \hat{y}_k)_{k\ge 1}$ such that they all satisfy:*

$$
\begin{aligned}
& y_k = \alpha_k v_{k-1} + (1-\alpha_k)x_{k-1}, \\
& \tilde{x}_k = T_{1/L_k}(y_k), \text{ with line search or backtracking.} \\
& v_k = x_{k-1} + \alpha_k^{-1}(\tilde{x}_k - x_{k-1}), \\
& \hat{y}_k = \operatorname{argmin}\left\{F(y) : y \in \{x_{k-1}, \tilde{x}_k\}\right\}, \\
& \eta_k \text{ s.t: } F(x_k) - F(\hat{y}_k) \le -1/(2\eta_k)\|\mathcal{G}_{1/\eta_k}(\hat{y}_k)\|^2, \eta_k \ge \eta_{k-1}, \\
& x_k = T_{1/\eta_k}(\hat{y}_k).
\end{aligned}
$$

The following theorem states the fact that the algorithm should eventually terminate if the objective function is bounded below.

{thm:nes-mono-wcnvx-convergence}

**Theorem 4.5.2 (convergence of Nesterov's monotone scheme nonconvex)**
*Suppose that the sequences $(y_{k+1}, v_k, x_k)_{k\ge 0}$ and $(\hat{y}_k, \tilde{x}_k)_{k\ge 0}$, $(\alpha_k)_{k\ge 0}$ satisfy Definition 4.5.1. Assume that $F$ is bounded below with $F^+ := \inf_x F(x)$. Then for all $N \ge 1$ it has*

$$
\min_{1\le k\le N} \|\mathcal{G}_{1/\eta_k}(\hat{y}_k)\|^2 \le \frac{2\overline{\eta}_N}{N}(F(x_{-1}) - F^+).
$$

*Here, $\overline{\eta}_k = \max_{i=0,\ldots,k} \eta_i$. If the monotone routine in Algorithm 4 is used, then it's bounded above by $2(q_g + L)$.*

*Proof.* $\overline{\eta}_k = \max_{i=0,\dots,k} \eta_i$ Using Lemma 4.1.6 it has from the descent condition of monotone routine that for all $k \geq 1$,

$$
\begin{aligned}
0 &\leq F(\hat{y}_k) - F\left(T_{1/\eta_k}\hat{y}_k\right) - \frac{1}{2\eta_k}\|\mathcal{G}_{1/\eta_k}(\hat{y}_k)\|^2 \\
&= \min(F(x_{k-1}), F(\tilde{x}_k)) - F(x_k) - \frac{1}{2\eta_k}\|\mathcal{G}_{1/\eta_k}(\hat{y}_k)\|^2 \\
&\leq F(x_{k-1}) - F(x_k) - \frac{1}{2\eta_k}\|\mathcal{G}_{1/\eta_k}(\hat{y}_k)\|^2 \\
&\leq F(x_{k-1}) - F(x_k) - \frac{1}{2\overline{\eta}_k}\|\mathcal{G}_{1/\eta_k}(\hat{y}_k)\|^2.
\end{aligned}
$$

Telescoping it has:

$$
\begin{aligned}
0 &\leq \left(\sum_{i=1}^{N} F(x_{i-1}) - F(x_i)\right) - \frac{1}{2\overline{\eta}_N}\sum_{i=1}^{N}\|\mathcal{G}_{1/\eta_i}(\hat{y}_k)\|^2 \\
&= F(x_0) - F(x_N) - \frac{1}{2\overline{\eta}_N}\sum_{i=1}^{N}\|\mathcal{G}_{1/\eta_i}(\hat{y}_k)\|^2 \\
&\leq F(x_0) - F(x_N) - \frac{N}{2\overline{\eta}_N}\left(\min_{1\leq i\leq N}\|\mathcal{G}_{1/\eta_i}(\hat{y}_i)\|^2\right) \\
&\leq F(x_0) - F^+ - \frac{N}{2\overline{\eta}_N}\left(\min_{1\leq i\leq N}\|\mathcal{G}_{1/\eta_i}(\hat{y}_i)\|^2\right) \\
&\leq F(x_{-1}) - F^+ - \frac{N}{2\overline{\eta}_N}\left(\min_{1\leq i\leq N}\|\mathcal{G}_{1/\eta_i}(\hat{y}_i)\|^2\right).
\end{aligned}
$$

Finally, we show $\overline{\eta}_k \leq 2(q_g - L)$. If there exists $k$ such that $\eta_k \geq q_g - L$ in the algorithm, then by Lemma 4.1.6 the condition $F(x_k) - F(\hat{y}_k) \leq -1/(2\eta_k)\|\mathcal{G}_{1/\eta_k}(\hat{y}_k)\|$ for all possible $\hat{y}_k \in \mathbb{R}^n$, therefore Algorithm 4 won't increase the value of $\eta_k$ in the future iteration. It has for all $i \geq k$, $\eta_i = \eta_k$. Suppose that some $\eta_i > 2(q_g + L), i \geq k$ then it means there exists $\eta_k > q_g + L$, this contradicts what we had right before, hence impossible and $\eta_i \leq 2(q_g + L)$ is an upper bound. ■

**Remark 4.5.3** The convergence claim still works for restarts.

A stronger result on the convergence rate of $\|\mathcal{G}_{1/\eta_k}(y_k)\|$ can be obtained if, we assume that the function $F = f + g$ satisfies Assumption 4.1.7. See Nesterov's book [10] for more details.

## 4.6 Restarting with function values for linear convergence

We adapted and improved prior theories on FISTA restart with global linear convergence from Alamo [1] for our AMAPG method. The following definition, gives the quadratic growth property of $f$ which allows for a fast linear convergence rate using adaptive restarts.

{ass:q-growth-ch2}

**Assumption 4.6.1 (quadratic growth condition)** Let $F = f + g$ satisfies Assumption 4.1.7 so that minimizers exists and, it's bounded below. Denote $F^+ = \inf_x F(x)$. Denote $X^+ = \operatorname*{argmin}_x F(x)$ and for all $x \in \mathbb{R}^n, \bar{x} \in \Pi_{X^+}x$ there exists $\mu > 0$ such that

$$F(x) - F^+ \geq \frac{\mu}{2}\|x - \bar{x}\|^2.$$

Let's introduce our first set of restart conditions which denote it by $\mathbf{E}_\chi^a$. $\mathbf{E}_\chi^a$ uses function values in the inner loop (Algorithm 5). Let $k$ denote the inner loop counter and define $m = \lfloor k/2 \rfloor + 1$, $\mathbf{E}_\chi^a$ is defined as:

{ineq:rxxapg-exit-cond}

$$\mathbf{E}_\chi^a \iff f(x_m) - f(x_k) \leq \exp(-1)(f(x_{-1}) - f(x_m)). \tag{4.6.1}$$

**Algorithm 7** Linear convergence restarted AMAPG

**Input:**

| | |
|---|---|
| $f : \mathbb{R}^n \to \mathbb{R}$ | Lipschitz Smooth |
| $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ | Weakly Convex |
| $x_{-1}$ | Vector |
| $M \in \mathbb{N}$ | Integer |
| $\epsilon \in \mathbb{R}$ | Number |
| $\mathbf{L}$ | Algorithm 1 or 2 |
| $\mathbf{M}$ | Algorithm 3 or 4 |
| $L := 1$ | $L > 0$ |
| $r := 0.5$ | $r \in (0, 1)$ |
| $\rho := 2^{1/1024}$ | $\rho \in (0, 1)$ |

$n_0 := 0; z_0 = x_{-1}$.
$z_1, p_0, \overline{L}_1, G^{(0)} := \mathbf{AMAPG}(f, g, x_{-1}, L, r_j, \rho, N_{\min} = n_0, N = M, \epsilon, \mathbf{L}, \mathbf{M}, \mathbf{E}^a_\chi)$.
$n_1 := p_0$.
$M := M - n_1$.
**for** $j = 1, 2, \ldots, M$ **do**
    **if** $M \leq 0$ **or** $G^{(j-1)} \leq \epsilon$ **then**
        **break**
    **end if**
{alg:rxxapg-iloop-done} $z_{j+1}, p_j, \overline{L}_{j+1}, G^{(j)} := \mathbf{AMAPG}(f, g, z_j, \overline{L}_j, r, \rho, N_{\min} = n_j, N = M, \epsilon, \mathbf{L}, \mathbf{M}, \mathbf{E}^a_\chi)$.
    $M := M - p_j$.
    $\overline{L}_{j+1} = \max(\overline{L}_j, \overline{L}_{j+1})$.
{alg:rxxapg-restart-if} **if** $f(z_j) - f(z_{j+1}) > \exp(-1)(f(z_{j-1}) - f(z_j))$ **then**
        $n_{j+1} := 2p_j$.
    **else**
        $n_{j+1} := p_j$.
    **end if**
{alg:rxxapg} **end for**

Algorithm 7 implements a restarted AMAPG with condition $\mathbf{E}^a_\chi$ stated in (4.6.1) and, it has a fast linear convergence rate. The following observation about it is crucial for deriving its convergence rate.

{obs:rxxapg}

**Observation 4.6.2** *If the outer loop runs for $j = 1, 2, \ldots, J$ iterations with $M \geq J$, so Algorithm 7 terminated due to $G^{(J-1)} \leq \epsilon$. Then, it has*

(i) *for all $J \geq j \geq 1$ it has $n_j \leq n_{j+1}$ hence $(n_j)_{j\geq 0}^J$ is monotone increasing;*

(ii) *for all $J - 1 \geq j \geq 1$ it has $p_{j-1} \leq p_j$ so $(p_j)_{j\geq 1}^J$ is monotone excluding $p_J$ obtained by the last iteration.*

Explanations for the observations now follow. For all $1 \leq j < J$, $n_j$ is passed in as the lower bound $N_{\min}$ for AMAPG inner loop (Algorithm 5), $p_j$ is the actual number of iteration by the inner loop therefore, it has $p_j \geq n_j$ because of Observation 4.3.1. $n_{j+1}$ is the minimum iteration required for the next execution of AMAPG inner loop, and it has $n_{j+1} = p_j$ if at line 13 is true otherwise, $n_{j+1} = 2p_j$. As a consequence, for $j < J$, it has either $p_{j+1} \geq n_{j+1} = 2p_j \geq 2n_j \geq n_j$ or $p_{j+1} \geq n_{j+1} = p_j \geq n_j$. Both $(n_j)_{j\geq 0}, (p_j)_{j\geq 0}$ are nondecreasing sequences.

When $j = J$ it's not necessarily true that $p_J \geq n_J$ because on the last iteration, inner loop can exit through condition $G_k^+ \leq \epsilon$ at line 17 of Algorithm 5, but $n_J \geq p_{J-1} \geq n_{J-1}$ still.

**Notations now follow.** Since the Algorithm 7 has a loop in variable $j$ and, an inner loop implemented by Algorithm 5 in variable $k$, we denote $x_k^{(j)}$ for the iterates $x_k$ in the inner loop during the $j$ iteration of the outer loop together. We make the convention $x_{-1}^{(j+1)} = z_{j+1} = x_{p_j}^{(j)}$ for consistencies across the theorems from previous sections.

For example at line 10 of Algorithm 7 at the $j$ iteration, the inner loop returns $x_{p_j}$ as the last iterate. So $x_{p_j}$ is assigned to $z_{j+1}$ by the outer loop, and it has $x_{p_j} = x_{p_j}^{(j)} = z_{j+1}$. It continues and $z_{j+1}$ will be the initial iterate pass into the inner loop for the $j+1$ iteration, and hence $z_{j+1} = x_{-1}^{(j+1)}$.

Let $e$ denotes the base of natural log. The following lemma will assert a lower bound on the number of iteration required to achieve a certain optimality on the objective function $F$ using the quadratic growth assumption.

{lemma:prog-ratio} **Lemma 4.6.3 (maximum iteration needed for an optimality gap ratio)**
*Suppose that $F = f + g$ satisfies Assumption 4.6.1 so $F^+ := \inf_x F(x)$, $X^+$ is the set of minimizers, and $\mu > 0$ is the quadratic growth constant. Let the sequence $(x_k)_{k\geq -1}$ be generated by AMAPG (Definition 5). Then it has*

$$\forall k \geq \left\lfloor \frac{2\sqrt{1+e}}{\sqrt{\mu \widehat{L}_k^{-1}}} \right\rfloor : \; F(x_k) - F^+ \leq e^{-1}(F(x_{-1}) - F(x_k)).$$

*Where $\widehat{L}_k$ is defined by:*

$$\widehat{L}_k := \max\left(L_0, \left(\frac{1}{k}\sum_{i=1}^{k}\sqrt{L_i^{-1}}\right)^{-2}\right).$$

*Proof.* For all $k \geq 0$, denote minimizer $x_k^+ = \Pi_{X^+} x_k$ so, $F(x_k^+) = F^+$. From Theorem 4.2.11

it has

$$F(x_k) - F^+ \le \frac{2\widehat{L}_k}{(2+k)^2}\|x_k - x_{-1}^+\|^2 \le \frac{4\widehat{L}_k}{\mu(2+k)^2}(F(x_{-1}) - F^+).$$

The second inequality comes from Assumption 4.6.1 directly. Suppose that $k \ge 2\sqrt{1+e}\left(\mu\widehat{L}_k^{-1}\right)^{-1/2}$. Denote $\gamma_k = 4\widehat{L}_k\mu^{-1}(2+k)^{-2}$. We make the following assumption first, and it will be proved later:

(a) It has $\gamma_k \in (0,1)$.

Using the above we have inequality:

$$\begin{aligned}
0 &\le \gamma_k(F(x_{-1}) - F^+) - (F(x_k) - F^+) \\
&= \gamma_k(F(x_{-1}) - F(x_k)) - (1-\gamma_k)(F(x_k) - F^+). \\
\underset{\text{(a)}}{\iff} F(x_k) - F^+ &\le \gamma_k(1-\gamma_k)^{-1}(F(x_{-1}) - F(x_k)).
\end{aligned}$$

Continuing it has

$$\begin{aligned}
F(x_k) - F^+ &\le \gamma_k(1-\gamma_k)^{-1}(F(x_{-1}) - F^+) \\
&= \frac{4\widehat{L}_k}{\mu(2+k)^2}\left(1 - \frac{4\widehat{L}_k}{\mu(2+k)^2}\right)^{-1}(F(x_{-1}) - F^+) \\
&= 4\widehat{L}_k(\mu(2+k)^2 - 4\widehat{L}_k)^{-1}(F(x_{-1}) - F^+) \\
&\le 4\widehat{L}_k\left(\mu\left(2 + \left\lfloor \frac{2\sqrt{1+e}}{\sqrt{\mu/\widehat{L}_k}}\right\rfloor\right)^2 - 4\widehat{L}_k\right)^{-1}(F(x_{-1}) - F^+) \\
&\le 4\widehat{L}_k\left(\mu\left(\frac{2\sqrt{1+e}}{\sqrt{\mu/\widehat{L}_k}}\right)^2 - 4\widehat{L}_k\right)^{-1}(F(x_{-1}) - F^+) \\
&= 4\widehat{L}_k\left(\mu\left(\frac{4(1+e)\widehat{L}_k}{\mu}\right) - 4\widehat{L}_k\right)^{-1}(F(x_{-1}) - F^+) \\
&= 4\widehat{L}_k\left(4\widehat{L}_k(1+e) - 4\widehat{L}_k\right)^{-1}(F(x_{-1}) - F^+) = e^{-1}(F(x_{-1}) - F^+).
\end{aligned}$$

**The proof for (a)** now follows. From the assumption on $k$ it has:

$$0 \ge \left\lfloor \frac{2\sqrt{1+e}}{\sqrt{\mu/\widehat{L}_k}}\right\rfloor - k > \frac{2\sqrt{1+e}}{\sqrt{\mu/\widehat{L}_k}} - 1 - k > \frac{2}{\sqrt{\mu/\widehat{L}_k}} - 1 - k$$

$$
\begin{aligned}
&> \frac{2}{\sqrt{\mu/\widehat{L}_k}} - (2+k) = (2+k)\left(\frac{2}{(k+2)\sqrt{\mu/\widehat{L}_k}} - 1\right) \\
&= (2+k)(\sqrt{\gamma_k} - 1).
\end{aligned}
$$

■

The following proposition places an upper bound on the estimated Lipschitz constant from the line search routine specified in Algorithm 1, 2. This is crucial to derive a global convergence properties of the algorithm based on the parameters of the objective function.

{prop:bnded-lip-ls}

**Proposition 4.6.4 (Lipschitz line search estimates are bounded)** *Suppose that $F = f + g$ satisfies Assumption 4.1.7. Choose such $F$ for Algorithm 5 so, it generates the sequence $(L_k)_{k\geq 0}$. Then, the sequence $(\widehat{L}_k)_{k\geq 1}$ from Theorem 4.2.11 is bounded above and, it has*

$$
\widehat{L}_k := \max\left(L_0, \left(\frac{1}{k}\sum_{i=1}^{k}\sqrt{L_i^{-1}}\right)^{-2}\right) \leq \overline{L} \leq \max\left(L_0, 2L\right).
$$

*Proof.* The following two facts are clear.

(i) A line search is triggered in Algorithm 1, 2 if and only if $L_{k+1} = 2L_k$ for some $k \geq 0$ and, it implies that there exists some $x \in \mathbb{R}^n, y \in \mathbb{R}^n$ such that $D_f(x, y) > L_k/2\|x - y\|^2$.

(ii) For all $k = 0, 1, \ldots$, if $L_k \geq L$, then it has $D_f(x, y) \leq L_k/2\|x - y\|^2$ for all $x, y \in \mathbb{R}^n$ which implies $L_{k+1} \leq L_k$ because the line search wasn't triggered.

For contradiction let's assume that there exists $k \geq 1$ such that a line search is triggered for $L_k > L$. From (i) it has $L_{k+1} = 2L_k > 2L$ so $L_{k+1} > L_k$, but (ii) showed that assumption $L_k > L$ implies $L_{k+1} \leq L_k$, which is a contradiction. Therefore, if $L_{k+1} = 2L_k$ then it must be that $L_k < L$ so, the highest value it can achieve is either $L_0$, or $2L$. ■

Continuing with the quadratic growth assumption, the following lemma states the result that $p_j$ in Algorithm 7 is bounded above and there exists a $j \geq 1$ such that $n_{j+1+k} = p_{j+k}$ for all $k \geq 0$ until it terminates.

{lemma:rxxapg-inner-bnds}

**Lemma 4.6.5 (inner iteration is bounded above)** *Suppose that $F = f + g$ satisfies Assumption 4.6.1, denote $F^+ := \inf_x F(x)$ and, $X^+$ as the set of minimizers. Consider any $j \geq 1$ iteration experienced by the outer loop. Define $\bar{p} := \frac{4\sqrt{2L(1+e)}}{\sqrt{\mu}}$, then $p_j \leq \bar{p}, n_j \leq \bar{p}$.*

*Proof.* The end result is constructed upon the following intermediate results that are proved at the end:

(i) If $k \geq \bar{p}$, then exit condition $\mathbf{E}_\chi^a$ is true hence $p_j \leq \max(\bar{p}, n_j)$ for all $j \geq 0$.

(ii) If $p_{j-1} \leq p_j \leq \bar{p}$ then $n_{j+1} \leq \bar{p}$.

Take note that $n_0 = 0, n_1 = p_0$ hence (i) gives $p_0 \leq \bar{p}$, and $p_1 \leq \max(\bar{p}, n_1) = \max(\bar{p}, p_0) \leq \bar{p}$. We now have the base case: $p_0 \leq p_1 = n_0 \leq \bar{p}$. Inductively assume $p_{j-1} \leq p_j \leq \bar{p}$ then:

$$p_{j+1} \underset{\text{(i)}}{\leq} \max(\bar{p}, n_{j+1}) \underset{\text{(ii)}}{\leq} \bar{p}.$$

Therefore, for all $j \geq 0$, $p_j \leq \bar{p}$, and $n_{j+1} \leq \bar{p}$.

**Proof of (i)**. Recall exit condition in (4.6.1) has $m = \lfloor k/2 \rfloor + 1$. Starting with the statement hypothesis it has $k \geq \bar{p}$ therefore:

$$\begin{aligned}
0 &\leq k/2 - \bar{p}/2 \leq \lfloor k/2 \rfloor + 1 - \bar{p}/2 = m - \bar{p}/2 \leq m - \lfloor \bar{p}/2 \rfloor \\
&= m - \left\lfloor \frac{2\sqrt{2L(1+e)}}{\sqrt{\mu}} \right\rfloor \leq m - \left\lfloor \frac{2\sqrt{\widehat{L}_k(1+e)}}{\sqrt{\mu}} \right\rfloor.
\end{aligned}$$

On the last inequality we used Proposition 4.6.4 which has $\widehat{L}_k \leq 2L$. Observe that the inequality allow us to apply Lemma 4.6.3 with $m = k$ which yields:

$$e^{-1} \geq \frac{F\left(x_m^{(j)}\right) - F^+}{F\left(x_{-1}^{(j)}\right) - F\left(x_m^{(j)}\right)} \geq \frac{F\left(x_m^{(j)}\right) - F\left(x_k^{(j)}\right)}{F\left(x_{-1}^{(j)}\right) - F\left(x_m^{(j)}\right)} \implies \mathbf{E}_\chi^a.$$

At line 17 of Algorithm 5, Since $\mathbf{E}_\chi^a$ is true, it will exit if $k \geq N_{\min} = n_j$ and, $G_k \leq \epsilon$ will only cause it to exit earlier therefore it has $p_j \leq \max(\bar{p}, n_j)$.

**Proof of (ii)**. Inductively assume that $p_{j-1} \leq p_j \leq \bar{p}$. If $p_{j-1} \leq \bar{p}/2$ then the if, else statement at line 13 in Algorithm 7 implies that $n_j \leq \max(p_{j-1}, 2p_{j-1}) \leq \bar{p}$. Otherwise, $p_{j-1} > \bar{p}/2$, and using Proposition 4.6.4 it means

$$p_{j-1} \geq \bar{p}/2 = \frac{2\sqrt{2L(1+e)}}{\sqrt{\mu}} \geq \left\lfloor \frac{2\sqrt{\widehat{L}_k(1+e)}}{\sqrt{\mu}} \right\rfloor.$$

The above inequality allows us to use Lemma 4.6.3 which yields

$$e^{-1} \geq \frac{F\left(x_{p_{j-1}}^{(j-1)}\right) - F^+}{F\left(x_{-1}^{(j-1)}\right) - F\left(z_{p_{j-1}}^{(j-1)}\right)} \geq \frac{F(z_j) - F(z_{j+1})}{F(z_{j-1}) - F(z_j)} \implies n_{j+1} = p_j.$$

This is true because, condition at line 13 in Algorithm 7 is never satisfied hence $n_{j+1} = p_j \leq \bar{p}$. ■

We just show that the sequence $n_j$ is bounded above, and it must have a limit because it's also non-decreasing from Observation 4.6.2, which implies that at some point, the doubling of $n_{j+1} = 2p_j$ must stop for the outer loop. The following lemma shows that when it happened, the restarted AMAPG (Algorithm 7) will always terminate after a finite number of iterations of the outer loop.

{lemma:rxxapg-outer-itr-bnd}

**Lemma 4.6.6 (bounds on outer iteration counts)** *Let $F = f + g$ satisfies Assumption 4.1.7. Suppose we apply Algorithm 7 on $F = f + g$. For all $\epsilon > 0$ used to terminate the algorithm, define $T_\epsilon = \lceil \ln(2\epsilon^{-2}(F(z_0) - F^+)) \rceil$. Assume that after iteration $j$, no doubling occurred in the if statement at line 13, i.e: $n_{t+1} = p_t$ for $t \geq j$, then it must terminate before, or at iteration $j + T_\epsilon$.*

*Proof.* Suppose that since the $j \geq 2$ th iteration, there is no period doubling for $T_\epsilon$ number of iterations in the outer loop of restarted AMAPG by Algorithm 7, i.e: $n_{t+1} = p_t$ for $j \leq t \leq j + T_\epsilon - 1$. Let's denote $s = j + T_\epsilon - 1$ for better notations, so for $t$ such that $j \leq t \leq s$, it has $n_{t+1} = p_t$.

Our goal now is to show that at iteration $j = s + 1$ of Algorithm 7 it must have $G^{(s)} \leq \epsilon$ making an exit due to $G_0 \leq \epsilon$ at line 6 of Algorithm 5. This is one of many ways the restarted AMAPG can exit, if we only focus on this condition it gives an upper bound on $T_\epsilon$.

Consider the start of the $s$ th iteration of the outer loop by Algorithm 13. Let's denote $L_0^{(s)}$ for the $L_0$ in the inner loop by Algorithm 5 at line 1; denote $G_0^{(s)}$ for the $G_0$ in the inner loop by Algorithm 5 at line 4. Then, it would give the following chain of inequalities

$$
\begin{aligned}
\frac{1}{2}\left(G_0^{(s)}\right)^2 &\underset{\text{(a)}}{=} \frac{1}{2}\left\|\sqrt{L_0^{(s)}}\left(z_s - T_{1/L_0^{(s)}}(z_s)\right)\right\|^2 \\
&\underset{\text{(b)}}{=} \frac{1}{2L_0^{(s)}}\left\|\mathcal{G}_{1/L_0^{(s)}}(z_s)\right\|^2 \\
&\underset{\text{(c)}}{\leq} F(z_s) - F\left(x_0^{(s)}\right) \\
&\underset{\text{(d)}}{\leq} F(z_s) - F\left(x_{p_s}^{(s)}\right) \\
&= F(z_s) - F(z_{s+1}) \\
&\underset{\text{(e)}}{\leq} \exp(-T_\epsilon)\left(F(z_{s-T_\epsilon}) - F(z_{s-T_\epsilon+1})\right) \\
&= \exp(-T_\epsilon)(F(z_{j-1}) - F(z_j)) \\
&\underset{\text{(f)}}{\leq} \exp(-T_\epsilon)(F(z_0) - F(z_j))
\end{aligned}
$$

$$
\begin{aligned}
&\underset{\text{(g)}}{\le} \left(\frac{2(F(z_0) - F^+)}{\epsilon^2}\right)^{-1} (F(z_0) - F^+) \\
&= \epsilon^2/2.
\end{aligned}
$$

(a) At the $s$ iteration of the outer loop by Algorithm 7, $z_s$ is passed into the inner loop by Algorithm 5 with $x_{-1} = z_s$. Therefore, at line 3 in Algorithm 5 it calls Armijo line search by Algorithm 1 with $x_{-1} = z_s, \alpha_0 = 1$ which means $y^+ = x_{-1} = z_s$ at line 3 and, $x^+ = T_{1/L^+}(z_s)$ at line 7. Coming back to line 3 in Algorithm 5, it assigns $y_0 = y^+ = z_s, x_0 = x^+ = T_{1/L^+}(z_s)$ and $L_0 = L^+$. Assuming the line search went successful, it will have $D_f(z_s, x_0) \le L_0^{(s)}/2\|z_s - x_0\|^2$.

(b) We used definition of gradient mapping in Definition 4.1.5.

(c) By the assumption that the line search in Algorithm 1 is successful at $z_s$ back in item (a), here we can use (4.2.2) in Theorem 4.2.7.

(d) AMAPG implemented via Algorithm 5 is monotone in function value for the use of $\mathbf{M}$ that is either Algorithm 3 or 4, so it has $F\left(x_{p_s}^{(s)}\right) \le F\left(x_0^{(s)}\right)$.

(e) Here we used the assumption that no doubling occurs so $n_{t+1} = p_t$ for all $j \le t \le s$ meaning that line 13 in Algorithm 7 has $F(z_t) - F(z_{t+1}) \ge e^{-1}(F(z_{t-1}) - F(z_t))$. Therefore, we can recursively unroll it for $T_\epsilon$ many iterations starting with $t = s = j + T_\epsilon - 1$ ending with $t = j$.

(f) We used the monotone property of subroutine $\mathbf{M}$ in AMAPG again so $F(z_0) \ge F(z_{j-1})$.

(g) We substituted $T_\epsilon = \lceil \ln(2\epsilon^{-2}(F(z_0) - F^+)) \rceil$ and, removing $\lceil\cdot\rceil$ to make for the $\le$ inequality. We also replaced $F(z_j)$ by $F^+$ the minimum which is always smaller.

Therefore, it has $G_0^{(s)} \le \epsilon$, hence it must have terminated at, or before iteration $s$. ■

{thm:rxxapg-total-itr-bnds}

**Theorem 4.6.7 (bounds on the total iterations)** *Let $F = f + g$ satisfies Assumption 4.6.1. Suppose that we applied Algorithm 7 to $F$. For any $\epsilon > 0$, assuming that it terminated at $j = J < M$ iteration. Then the total number of iterations has an upper bound:*

$$
\sum_{i=0}^{J-1} p_i \le \frac{8\sqrt{2L(1+e)}}{\sqrt{\mu}} \left\lceil \ln\left(\frac{2(F(z_0) - F^+)}{\epsilon^2}\right)\right\rceil .
$$

*Proof.* Firstly, Algorithm 7 must terminate within finite many iterations under Assumption 4.6.1 for all total budget $M \in \mathbb{N}$. This is because Observation 4.6.2 shows that $n_j$ is a non-decreasing sequence, Lemma 4.6.5 shows that $n_j \le \bar{p}$ under the quadratic growth assumption, therefore $n_j$ must converge which implies that doubling of $n_{j+1} = 2p_j$ must stop. Finally, since the doubling must stop at some $j$, Lemma 4.6.6 applies hence, it terminates at most $j + T_\epsilon$ iteration.

Using it let's assume that it executed for $j = 1, 2, \cdots, J$ and $M$ is large enough to achieve optimality $G^{(J)} \leq \epsilon$ right at the start of iteration $J+1$. This is one of the way the algorithm can terminate, making it a sufficient condition for deriving the upper bound on the total number of iterations.

Let's represent $J$ by: $J = m + nT_\epsilon$ with $0 \leq m < T_\epsilon$. The following intermediate results are important to the proof.

(a) $n_{J-lT_\epsilon} \leq n_{J-lT_\epsilon} \leq (1/2)^l n_J$ for all $l = 1, \dots, n$. This is true because Lemma 4.6.6 showed that doubling must have occurred within a period of $T_\epsilon$ at least once.

(b) $n_j \leq n_{j+1}$ for all $0 \leq j \leq J-1$. The sequence is monotone non decreasing from Observation 4.6.2. In addition, $0 \leq m < T_\epsilon$ by assumption.

(c) $n_j \leq \bar{p}$ with $\bar{p} = 4\sqrt{2L(1+e)}/\sqrt{\mu}$, which is proved in Lemma 4.6.5.

The upper bound on the total number of iterations of AMAPG over $J$ iteration of outer loop is given by:

$$
\begin{aligned}
\sum_{i=0}^{J} n_i &= \sum_{i=0}^{m+nT_\epsilon} n_i \\
&= \sum_{i=0}^{m} n_i + \sum_{l=0}^{n-1} \sum_{i=1}^{T_\epsilon} n_{m+i+lT_\epsilon} \\
&\underset{(b)}{\leq} T_\epsilon n_m + \sum_{l=0}^{n-1} T_\epsilon n_{m+(l+1)T_\epsilon} \\
&= T_\epsilon \sum_{l=0}^{n} n_{m+lT_\epsilon} = T_\epsilon \sum_{l=0}^{n} n_{J-lT_\epsilon} \\
&\underset{(a)}{\leq} T_\epsilon \sum_{l=0}^{n} (1/2)^l n_J \leq T_\epsilon \sum_{l=0}^{\infty} (1/2)^l n_J \\
&\leq 2T_\epsilon n_J \underset{(c)}{\leq} 2T_\epsilon \bar{p}.
\end{aligned}
$$

The total number of iterations is bounded by:

$$
\sum_{i=0}^{J-1} p_i \leq \sum_{i=0}^{J} n_i \leq \frac{8\sqrt{2L(1+e)}}{\sqrt{\mu}} \left\lceil \ln\left(\frac{2(F(z_0) - F^+)}{\epsilon^2}\right) \right\rceil .
$$

Note that $n_0 = 0$. ■

The final results from above theorem provide the convergence rate of iterates and the complexity of restart AMAPG under Assumption 4.6.1. We denote $\kappa := L/\mu$.

{thm:rxxapg-cnvg-complexity} **Theorem 4.6.8 (restarted AMAPG convergence and complexity)**
*Let $F = f + g$ satisfy Assumption 4.6.1, and apply it to Algorithm 7. Let $J$ be the total number of iteration performed to achieve accuracy in the outer loop. For each $0j \le J$, $p_j$ is the number of inner iterations of the inner loop. Let $K := \sum_{i=0}^{J-1} p_i$ be the total number of iterations of the inner loop. Let $z_j$ be the iterates returned by the inner loop. Then, the maximum $K$ needed to achieve optimality $\|z_J - z_J^+\| \le \delta$ is bounded by $\mathcal{O}\left(\frac{1}{\sqrt{\kappa}} \ln\left(\frac{1}{\kappa\delta}\right)\right)$.*

*Proof.* Suppose a total of $K := \sum_{i=0}^{J-1} p_i$ iteration were performed and, at line 10 of Algorithm 7 at iteration $j = J$ it achieved $G_0 \le \epsilon$ at line 5 of Algorithm 5. That will cause the inner AMAPG to return $G_0$ so, at line 10 it has $G^{(J)} < \epsilon$ in Algorithm 7. This is one of the ways the algorithm can terminate hence, it suffices for deriving an upper bound.

Now we show the convergence rate of $G^{(J)}$, let $k > 0$ and let $\epsilon = \sqrt{2}(F(z_0) - F^+)^{1/2}\exp(-k+1)$ then:

$$\ln\left(\frac{2(F(z_0) - F^+)}{\epsilon^2}\right) = 2(k-1).$$

Then it has from Theorem 4.6.7 that:

$$\begin{aligned}
&0 \le \frac{8\sqrt{2L(1+e)}}{\sqrt{\mu}}\left\lceil \ln\left(\frac{2(F(z_0) - F^+)}{\epsilon^2}\right)\right\rceil - K = \frac{8\sqrt{2L(1+e)}}{\sqrt{\mu}}\lceil 2(k-1)\rceil - K \\
&\le \frac{8\sqrt{2L(1+e)}}{\sqrt{\mu}}(2(k-1)+1) - K \\
\implies &0 \le k - 1 + 1/2 - \frac{K\sqrt{\mu}}{16\sqrt{2L(1+e)}} \\
&\le k - \frac{K\sqrt{\mu}}{16\sqrt{2L(1+e)}}.
\end{aligned}$$

This gives us

{ineq:rxxapg-cnvg-complexity-proof-p1}
$$\begin{aligned}
G^{(J)} &\le \epsilon = \sqrt{2}(F(z_0) - F^+)^{1/2}\exp(-k+1) \\
&= e\sqrt{2}(F(z_0) - F^+)^{1/2}\exp(-k) \\
&\le e\sqrt{2}(F(z_0) - F^+)^{1/2}\exp\left(-\frac{K\sqrt{\mu}}{16\sqrt{2L(1+e)}}\right) \\
&= e\sqrt{2}(F(z_0) - F^+)^{1/2}\exp\left(-\frac{K\sqrt{\kappa}}{16\sqrt{2+2e}}\right).
\end{aligned} \tag{4.6.2}$$

The above inequality shows a linear convergence rate of the quantity $G_0^{(J)}$ with respect to the total number of iterations required for AMAPG. Denote $z_J^+ = \Pi_{X^+} z_J$ then it has

{ineq:rxxapg-cnvg-complexity-proof-p2}

$$
\begin{aligned}
G^{(J)} &\underset{\text{(a)}}{=} \left\Vert \sqrt{L_0^{(J)}}\left(z_J - x_0^{(J)}\right)\right\Vert = \frac{1}{\sqrt{L_0^{(J)}}}\left\Vert L_0^{(J)}\left(z_J - x_0^{(J)}\right)\right\Vert \underset{\text{(a)}}{=} \frac{1}{\sqrt{L_0^{(J)}}}\left\Vert \mathcal{G}_{1/L_0^{(J)}}(z_J)\right\Vert \\
&\underset{\text{(b)}}{\ge} \frac{1}{\sqrt{2L}}\left\Vert \mathcal{G}_{1/L_0^{(J)}}(z_J)\right\Vert \\
&\underset{\text{(c)}}{\ge} \frac{L(\sqrt{L(\mu+L)} - L)}{\sqrt{L(\mu+L)}}\frac{1}{\sqrt{2L}}\left\Vert z_J - z_J^+\right\Vert \\
&= \frac{\sqrt{L}}{\sqrt{2}}\left(1 - \frac{1}{\sqrt{1+\mu/L}}\right)\left\Vert z_J - z_J^+\right\Vert \\
&\underset{\text{(d)}}{\ge} \frac{\mu\sqrt{L}}{\sqrt{2}}\left(\frac{\sqrt{2}-1}{\sqrt{2}}\right)\left\Vert z_J - z_J^+\right\Vert = \frac{\kappa}{\sqrt{L}}\left(\frac{\sqrt{2}-1}{2}\right)\left\Vert z_J - z_J^+\right\Vert .
\end{aligned}
\tag{4.6.3}
$$

(a) For the first equality, we assumed that $G^{(J)} < \epsilon$ in the outer loop and, it's returned by Algorithm 5 at line 5. Therefore, it has $G_0 = G^{(J)} = \sqrt{L_0^{(J)}}\left\Vert z^{(J)} - x_0^{(J)}\right\Vert$ because it calls Algorithm 1 (Armijo line search) with $\alpha_0 = 0$. The second equality comes by using Definition 4.1.5.

(b) Proposition 4.6.4 has $L_0^{(J)} \le 2L$.

(c) Using Theorem 2.1.14 from previous chapter and take note that Assumption 4.6.1 is Definition 2.1.3(iv).

(d) Define function $h = x \mapsto 1 - (1 + x/L)^{-1/2}$ is concave hence for all $x \in [0, L]$ it has:

$$
\begin{aligned}
1 - (1+\mu/L)^{-1/2} &= h(0(1-\mu) + \mu L) \\
&\ge 0(1-\mu) + \mu\left(1 - (1+1)^{-1/2}\right) \\
&= \mu\left(1 - \frac{1}{\sqrt{2}}\right) = \frac{\mu(\sqrt{2}-1)}{\sqrt{2}}.
\end{aligned}
$$

Using results of (4.6.2), (4.6.3) it has

$$
\left\Vert z_J - z_J^+\right\Vert \le \frac{2e\sqrt{2L}(F(z_0) - F^+)^{1/2}}{\kappa(\sqrt{2}-1)}\exp\left(-\frac{K\sqrt{\kappa}}{16\sqrt{2}+2e}\right).
$$

$\Vert z_J - z_J^+\Vert \le \delta$ is achieved by substituting $K$:

$$
K = \left(\frac{16\sqrt{2}+2e}{\sqrt{\kappa}}\right)\ln\left(\frac{2e\sqrt{2L}(F(z_0) - F^+)^{1/2}}{\kappa(\sqrt{2}-1)\delta}\right).
$$

This is an upper bound for the number of required iterations. When constants are ignored, it's $K \le \mathcal{O}\left(\frac{1}{\sqrt{\kappa}}\ln\left(\frac{1}{\kappa\delta}\right)\right)$. ■

**Remark 4.6.9** The convergence rate can be simplified a bit. On a calculator it has $(2e\sqrt{2})/(\sqrt{2}-1) < 19$ and, $1/(16\sqrt{2+2e}) > 0.025$ so

$$\|z_J - z_J^+\| \leq \frac{19\sqrt{L}(F(z_0) - F^+)^{1/2}}{\kappa} \exp\left(-\frac{K\sqrt{\kappa}}{40}\right).$$

## 4.7 Hoffman error bound and quadratic growth

The Hoffman error bound conditions is essential to analyzing optimization problem arises in applications with polytoptic constraints.

## 4.8 Numerical experiments

# Chapter 5

# Enhanced Primal Dual Methods for LP

# Bibliography


[1] T. Alamo, P. Krupa, and D. Limon, *Restart FISTA with global linear convergence*, Dec. 2019.

[2] J.-F. Aujol, L. Calatroni, C. Dossal, H. Labarrière, and A. Rondepierre, *Parameter-Free FISTA by adaptive restart and backtracking*, SIAM Journal on Optimization, (2024), pp. 3259–3285.

[3] A. Beck and M. Teboulle, *Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems*, IEEE Transactions on Image Processing, 18 (2009), pp. 2419–2434.

[4] J. V. Burke and P. Tseng, *A unified analysis of Hoffman's Bound via Fenchel Duality*, SIAM Journal on Optimization, 6 (1996), pp. 265–282.

[5] L. Calatroni and A. Chambolle, *Backtracking strategies for accelerated descent methods with smooth composite objectives*, SIAM Journal on Optimization, 29 (2019), pp. 1772–1798.

[6] O. Güler, *New proximal point algorithms for convex minimization*, SIAM Journal on Optimization, 2 (1992), pp. 649–664.

[7] O. Güler, A. J. Hoffman, and U. G. Rothblum, *Approximations to solutions to systems of linear inequalities*, SIAM Journal on Matrix Analysis and Applications, 16 (1995), pp. 688–696.

[8] A. Hoffman, *On approximate solutions of systems of linear inequalities*, Journal of Research of the National Bureau of Standards, 49 (1952), p. 263.

[9] 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.

[10] Y. Nesterov, *Lectures on Convex Optimization*, vol. 137 of Springer Optimization and Its Applications, Springer International Publishing, 2018.