Reading Notes

Alto

Last Compiled: May 8, 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 $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a differentiable function. Define Bregman Divergence:*

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

Assumption 1.0.2 (smooth plus nonsmooth) Let $F = f + g$ where $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ 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-wcnvx-generic} **Theorem 1.0.4 (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:*

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

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

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

■

{thm:cnvx-pg-ineq}

Theorem 1.0.5 (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 $\forall z \in \mathbb{R}^n$:

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

■

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.

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. [1].

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 Π_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. The gradient mapping in this case is:

{def:necoara-scnvx}

$$\mathcal{G}_{L_f} x = L_f(x - \Pi_X x).$$

Definition 2.1.2 (strong convexity) *Suppose f satisfies Assumption 2.1.1. Then $f \in \mathbb{S}(L_f, \kappa_f, X)$ is strongly convex iff*

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

Then it's not hard to imagine the following natural relaxation of the above conditions.

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 $x \in X$, $\bar{x} = \Pi_{X^+} x$. We define the following:

{def:neocara-qscnvx}

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

{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 $\mathbb{U}(L_f, \kappa_f, X)$.*

{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 $\mathbb{G}(L_f, \kappa_f, X)$.*

{def:necoara-qfg}

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

{def:necoara-peb}

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

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

{thm:qscnvx-means-qua}

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

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

$$\mathbb{S}'(L_f, \kappa_f, X) \subseteq \mathbb{U}(L_f, \kappa_f, X).$$

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

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

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}

$$\begin{aligned} f^+ - f(x_\tau) &\geq \langle \nabla f(x_\tau), \Pi_{X^+} x_\tau - x_\tau \rangle + \kappa_f/2\|x_\tau - \Pi_{X^+} x_\tau\|^2 \\ &= \langle \nabla f(x_\tau), \bar{x} - x_\tau \rangle + \kappa_f/2\|x_\tau - \bar{x}\|^2 \\ \iff \langle \nabla f(x_\tau), x_\tau - \bar{x} \rangle &\geq f(x_\tau) - f^+ + \kappa_f/2\|x_\tau - \bar{x}\|^2. \end{aligned} \tag{2.1.3}$$

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

(i) The inequality (2.1.3).

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

(iii) The inductive hypothesis with $k \geq 0$.

(iv) $\bar{x} = \Pi_{X^+} x$, 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:

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

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 $\kappa_f^{(n)} \ge 0$, 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*

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

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. [1]. 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*

$$\mathbb{S}'(L_f, \kappa_f, X) \subseteq \mathbb{G}(L_f, \kappa_f, X).$$

Proof. If $f \in \mathbb{S}'(L_f, \kappa_f, X)$ then Theorem 2.1.5 has $f \in \mathbb{U}(L_f, \kappa_f, X)$. 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. [1], right after it claims $\mathbb{U}(L_f, \kappa_f, X) \subseteq \mathbb{G}(L_f, \kappa_f/2, X)$ under convexity.

{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.

$$\begin{aligned}
\|x - \Pi_{X^+}x\| &\le \|x - \Pi_{X^+}x^+\| && (2.1.4)\\
&\le \|x - x^+\| + \|x^+ - \Pi_{X^+}x^+\| && (2.1.5)\\
&\le \|x - x^+\| + \beta\|x - \Pi_{X^+}x\| && (2.1.6)\\
\iff 0 &\le \|x - x^+\| - (1-\beta)\|x - \Pi_{X^+}x\|. && (2.1.7)
\end{aligned}$$

x^+ has descent lemma hence we have

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

Hence, it gives the quadratic growth condition. ■

Remark 2.1.12 It's unclear where convexity is used. However, it' still assumed in Necoara et al. paper.

Before we start, we will specialize Theorem 1.0.5 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}
$$\begin{aligned}
0 &\le f(z) - f(x^+) + \frac{L_f}{2}\|z - x\|^2 - \frac{L_f}{2}\|z - x^+\|^2\\
&= f(z) - f(x^+) + L_f\langle z - x^+, x^+ - x\rangle + \frac{L_f}{2}\|x - x^+\|^2.
\end{aligned} \tag{2.1.8}$$

Take note that when $z = x$ it has

{ineq:proj-grad2}
$$0 \le f(x) - f(x^+) - \frac{L_f}{2}\|x - x^+\|^2. \tag{2.1.9}$$

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 Assumtion 2.1.1. Suppose that $f \in \mathbb{F}(L_f, \mu_f, X)$ so it satisfies the quadratic function growth condition. For all $x \in \mathbb{R}^n$, define $x^+ = \Pi_X(x - L_f^{-1}\nabla f(x))$, definte projections onto the set of minimizers $x_\Pi^+ = \Pi_{X^+}x^+, X_\Pi = \Pi_{X^+}x$, then*

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

Proof. Using convexity, consider (2.1.8) with $z = x_\Pi^+$ it yields:

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

From the last line, it's can be equivalently expressed as:

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

{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:*

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

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

$$\begin{aligned}
\|x - x_\Pi\| &\le \|x - x_\Pi^+\| \le \|x - x^+\| + \|x^+ - x_\Pi^+\| \\
&= \frac{1}{L_f}\|L_f(x - x^+)\| + \|x^+ - x_\Pi^+\|
\end{aligned}$$

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

$$\iff \|x^+ - x_\Pi^+\| \ge \|x - x_\Pi\| - \frac{1}{L_f}\|L_f(x - x^+)\|. \tag{2.1.10}$$

Before we start, we list intermediate results and conditions which are going to be used in the proof that follows for the ease of referencing.

(i) The inequality (2.1.10). It uses the property of projection onto a set hence convexity of X^+ is not needed.

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

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

Skipping some algebra, the fraction simplifies to

$$
\frac{\kappa_f}{\kappa_f/L_f + 1 + \sqrt{\kappa_k/L_f + 1}}.
$$

This gives PEB condition. **We now show PEB implies QFG**. From the error bound condition using κ_f it has

$$
\kappa_f^2\|x - \bar{x}\|^2 \le \|\mathcal{G}_{L_f}(x)\|^2 \underset{(2.1.9)}{\le} 2L_f(f(x) - f(x^+)) \le 2L_f(f(x) - f^+).
$$

■

{thm:q-cnvx-hierarchy} The following theorem summarizes the hierarchy of the conditions listed in Definition 2.1.3.

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

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

Proof. $\mathbb{S}' \subseteq \mathbb{G}$ is proved in Theorem 2.1.9 and $\mathbb{G} \subseteq \mathbb{U}$ is proved in 2.1.7. $\mathbb{S} \subseteq \mathbb{S}'$ is obvious and it remains to show $\mathbb{U} \subseteq \mathbb{F}$. Let $f \in \mathbb{U}(\kappa_f, L_f, X)$, 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. [1].

2.1.3 Hoffman error bound and Q-SCNVX

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 2.1.17 (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}

$$(\forall x \in X)\ \langle x_{k+1} - (x_k + \alpha_k \nabla f(x_k)), x_{k+1} - x\rangle \leq 0. \tag{2.1.11}$$

Theorem 2.1.18 *feasible descent linear convergence under Q-SCNVX Under Assumption 2.1.1, assume that* f *is Q-CNVX with* μ_f, L_f*, then the sequence that satisfies Definition 2.1.17 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 (2.1.11), 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-CNVX 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.

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

Therefore, it has

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

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

2.1.5 Application, KKT of linear programming

This section extends and ideas in the discussion section of Necoara et al. [1].

Let X_1, X_2, Y be Hilbert spaces. Define linear mapping $E : X_1 \times X_2 \to Y := (x_1, x_2) \mapsto E_1 x_1 + E_2 x_2$ 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}
$$\inf_{x \in X_1 \times X_2} \left\{ \langle -c, x\rangle \mid Ex = b, x \in \mathcal{K} \times X_2 \right\}. \tag{2.1.12}$$

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 (2.1.12) 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 convex feasibility problem

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

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

$$\min_{v\in\mathbf{K}} \frac{1}{2}\|Ax-d\|^2. \tag{2.1.13}$$

It is minimizing a quadratic problem on a simple cone. Solving (2.1.12) can be approached by optimizing (2.1.13). 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

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

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

Let $f(v) = (1/2)\|Av - d\|^2$ to be the objective function of optimization problem (2.1.13). Its gradient, objective value, and Bregman Divergence have:

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

The value $\nabla f(v), f(v)$ when evaluated together, require minimal additional computations. 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 problem, the computation overhead for $f(u) - f(v), D_f(u, v)$ is very little if $\nabla f(u), \nabla f(v)$ is known.

Chapter 3

Advanced Enhancement Techniques in Accelerated Proximal Gradient

We review advanced enhancement techniques in Accelerated Proximal Gradient method. The review will be based on several papers.

There are several notable enhancements of the FISTA for function that are not strongly convex. Monotone variants of FISTA proposed by Beck and Nesterov imposes monotonicity in function value at the iterates. Backtracking strategies from Chambolle shows that the underestimating Lipschitz constant using a backtracking technique to choose a next iterate improves the average runtime of the algorithm in practice. They showed that the convergence rate is bounded by the estimates of the Lipschitz constant. Restart is a technique pioneer early by ??? . Necoara et al. [1] showed that there exists an optimal restarting interval to achieve fast line convergence rate for all functions with quadratic growth condition.

[?]

[?]

[?]

[?]

In this chapter, we will go through the details of these enhancements of FISTA and discuss why they are important in theories, and in practice.

3.1 Preliminaries

This section introduces the full scope of the theories in our analysis.

3.1.1 smooth plus nonsmooth weakly convex

{ass:sum-of-wcnvx} **Assumption 3.1.1 (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 L Lipschitz smooth and q_g weakly convex.

{def:wcnvx-fxn} **Definition 3.1.2 (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 $q \in \mathbb{R}$ such that the function $F + q/2\|\cdot\|^2$ is a convex function and q is the infimum of all such possible parameters.

Definition 3.1.3 (gradient mapping) *Suppose $F = f + g$ satisfies Assumption 3.1.1, 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 3.1.4 (weakly convex monotone descent)**
Let $F = f + g$ satisfies Assumption 3.1.1. Let $\bar{x} = T_{\beta,f,g}(x)$. Then, for all $x \in \mathbb{R}^n$, it has the following inequality:

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

And descent is possible when $\beta \ge (L + q)/2$. The maximum amount of descent is achieved when $\beta = L + q$.

Proof. Use Theorem 1.0.4. ■

3.1.2 smooth plus nonsmooth convex

{ass:standard-fista} The following assumption is strictly stronger than 3.1.1.

Assumption 3.1.5 (convex smooth and nonsmooth) Let $F = f + g$ where $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is L Lipschitz smooth, g is convex. Suppose that $\operatorname*{argmin}_{x \in \mathbb{R}^n} F(x) \neq \emptyset$.

Lemma 3.1.6 (proximal gradient inequality) *If $F = f + g$ satisfies Assumption 3.1.5, 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 1.0.5. ■

3.2 FISTA made simple

We make the proofs for FISTA with common enhancement technique available in simple proofs. We showcase the theories using a generic similar triangle representations of the algorithm which tremendously simplifies the arguments.

The following definition can capture monotone variants of FISTA with line search, including backtracking strategies.

"MAPG" stands for monotone accelerated gradient. We refer to "Generic Monotone Accelerated Proximal Gradient with line search" as "GMAPG".

{def:generic-mapg-ls}

Definition 3.2.1 (GMAPG)
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:

$$
\begin{aligned}
&y_k = \alpha_k v_{k-1} + (1-\alpha_k)x_{k-1},\\
&\tilde{x}_k = T_{L_k^{-1}}(y_k),\\
&v_k = x_{k-1} + \alpha_k^{-1}(\tilde{x}_k - x_{k-1}),\\
&D_f(\tilde{x}_k, y_k) \le \frac{L_k}{2}\|\tilde{x}_k - y_k\|^2,\\
&\textit{Choose any } x_k : F(x_k) \le \min(F(\tilde{x}_k), F(x_{k-1})).
\end{aligned}
$$

{lemma:apg-iterates}

Lemma 3.2.2 (accerated proximal gradient iterates relation)
The iterates $(x_k, v_k, y_k)_{k\ge 1}$ generated by Definition 3.2.1. 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:gmapg-ls-convergence} **Theorem 3.2.3 (generic GMAPG convergence)**
Let $F = f + g$ satisfy Assumptions 3.1.5. Let $(\alpha_k)_{k\ge 0}$ be a sequence such that $\alpha_k \in (0,1)$ for all $k \ge 1$ and $\alpha_0 \in (0,1]$. Let $\rho_k = (1-\alpha_{k+1})^{-1}\alpha_{k+1}^2\alpha_k^{-2}$ for all $k \ge 0$. Then, for all $x^+ \in \mathbb{R}^n, k \ge 1$, the convergence rate of GMAPG-LS (Definition 3.2.1) is given by:

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

If in addition, the algorithm is initialized with $L_0 \ge L$, $\alpha_0 = 1, x_0 = v_0 = T_{L_0}x_{-1} \in dom\ F$ and x^+ is a minimizer of F. Then, the convergence rate simplifies:

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

Proof. Define $z_k = \alpha_k x^+ + (1-\alpha_k)x_{k-1}$ for all $k \ge 1$. In the proof follows, the follow facts will be used. We list them in advance, and they will be labeled during the proof.

(i) Lemma 3.2.2.

(ii) The sequence $(\alpha_k)_{k\ge 1}$ has for all $k \ge 1$, $1 - \alpha_k = \alpha_k^2\alpha_{k-1}^2\rho_{k-1}$, $\alpha_k \in (0,1)$.

(iii) F is convex and hence $F(z_k) \le \alpha_k F(x^+) + (1-\alpha_k)F(x_{k-1})$.

(iv) $F(x_k) \le F(\tilde{x}_k)$ which is true by definition of GMAPG.

Now, using Theorem 1.0.5, it has for all $k \in \mathbb{N}$:

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

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

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

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

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 1.0.5 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 3.2.4 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:gmapg-generic-gradmap-convergence} **Theorem 3.2.5 (generic GMAPG gradient mapping convergence)** *Suppose that $F = f + g$ satisfies Assumption 3.1.5. Let the sequences (x_k, y_k, v_k) satisfy GMAPG (Definition 3.2.1). If in addition, the sequence $(\alpha_k)_{k\ge 0}$ has $\alpha_0 = 1$ and, GMAPG is initialized with $L_0 \ge L$, $v_0 = x_0 = T_{1/L_0,f,g}(x_{-1})$ for any $x_{-1} \in \mathbb{R}^n$ and there exists x^+ which is a minimizer of F. Then, the convergence of gradient mapping $\|\mathcal{G}_{1/L_k}(y_k)\|$ is described by the sequence $(\beta)_{k\ge 0}$ where $\beta_0 = 1$ and for all $k \ge 1$:*

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

It satisfies for all $k \ge 1$ the inequality:

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

Proof. The following intermediate results will be used in the proof. We stated prior to the proof for a sleeker exposition.

(a) From Definition 3.2.1, 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. 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 the sequence $(\beta_k)_{k\ge 0}$ stated in the statement hypothesis.

From these intermediate results, the convergence can be derived. From (a) it has for all $k \geq 0$:

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

Now, **let's proof intermediate results (a)**. From the definition of GMAPG it has

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

Using the above, and definition of GMAPG, it yields

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

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 the proximal gradient inequality with x^+ as a minimizer it yields:

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

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 3.2.3 that

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

■

{lemma:gmapg-seq-bnd} **Lemma 3.2.6 (specialized GMAPG momentum sequence)**
Suppose that the sequence $(\alpha_k)_{k\ge 0}$ has $\alpha_0 \in (0,1]$ and for all $k \ge 1$, it satisfies $\alpha_{k-1}^2\rho_{k-1}(1 - \alpha_k) = \alpha_k^2$. Suppose that $(\beta_k)_{k\ge 0}$ is a sequence such that $\beta_0 = 1$ and for all $k \ge 1$ it's defined as:

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

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 $k \ge 1$ it 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) $(\beta_k)_{k\ge 0}$ is monotone decreasing and it's strictly larger than zero.

(b) Because $\rho_k L_{k+1} L_k = 1$ 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:

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

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

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

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 $\alpha_k \in (0,1)$, β_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 β_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 3.2.7 A simpler upper bound is more useful in practice. For all $k \ge 1$ let $\hat{L}_k = \max_{i=0,1,\ldots,k} L_i$ then

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

This simplifies the convergence claim back in Theorem 3.2.3. The above inequality would work the same if we set:

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

{thm:gmapg-specialized-convergence}

Theorem 3.2.8 (specialized GMAPG convergence rate) *Suppose that $F = f + g$ satisfy Assumption 3.1.5. Let the sequence (x_k, v_k, v_k) satisfy GMAPG, definition 3.2.1 and, assume that the GMAPG 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*

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

Then, we have convergence claim:

(i)

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

(ii)

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

Proof. To see (i), use 3.2.6 and the remark with Theorem 3.2.3 because the assumptions of $x^+, (\alpha_k)_{k\ge 0}, (\rho_k)_{k\ge 0}$ is the same. To see (ii), the convergence claim from 3.2.5 simplifies with $\hat{L}_k \ge L_0$ and, it has

$$\begin{aligned}
\|\mathcal{G}_{1/L_k}(y_k)\| &\le \left(\frac{2\sqrt{\hat{L}_k L_0}L_k}{2+k}\right)\left(1 + \min(\rho_{k-1}, L_k^{-1}L_{k-1})^{1/2}\right)\|x^+ - v_0\| \\
&= \left(\frac{2\sqrt{\hat{L}_k L_0}L_k}{2+k}\right)\left(1 + L_k^{-1/2}L_{k-1}^{1/2}\right)\|x^+ - v_0\| \\
&\le \left(\frac{2\hat{L}_k L_k}{2+k}\right)\left(1 + L_k^{-1/2}L_{k-1}^{1/2}\right)\|x^+ - v_0\|.
\end{aligned}$$

■

3.3 Algorithmic description of GMAPG

There are several components to the GMAPG algorithm. This section will introduce various type of implementations that can be fitted into GMAPG in Definition 3.2.1.

{alg:armijo-ls}

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 function with Proximal operator
$x \in \mathbb{R}^n$	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

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

{alg:chambolle-btls}

Algorithm 2 Chambolle's Backtracking

Function ChamBT input:

$f : \mathbb{R}^n \to \mathbb{R}$	Convex Lipschitz smooth
$g : \mathbb{R}^n \to \overline{\mathbb{R}}$	Convex with Proximal Operator
$x \in \mathbb{R}^n$	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
Return: y^+, x^+, α^+, L^+

{alg:beck-mono}

Algorithm 3 Beck's monotone routine

Function BeckMono Inputs:

$f : \mathbb{R}^n \to \mathbb{R}$	
$g : \mathbb{R}^n \to \overline{\mathbb{R}}$	
$\tilde{x} \in \mathbb{R}^n$	
$x \in \mathbb{R}^n$	
ρ	Number $\rho \in (0, 1)$

$x^+ = \operatorname{argmin}\{(f + g)(z) : z \in \{\tilde{x}, x\}\}$.
Return: x^+, η

Algorithm 4 Nesterov's monotone routine

1: **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 $\eta > 0$

2: $\hat{y} := \operatorname{argmin}\{(f+g)(z), x \in \{\tilde{x}, x\}\}$.
3: $x^+ := T_{1/\eta}(\hat{y})$.
4: **while** $F(x^+) - F(\hat{y}) > -1/(2\eta)\|\mathcal{G}_{1/\eta}(\hat{y})\|^2$ **do**
5: $\quad \eta := 2\eta$.
6: **end while**
{alg:nes-mono} 7: **return:** x^+, η

Algorithm 5 GMAPG with Chambolle's backtracking

1: **Function GMAPG Inputs:**

$f : \mathbb{R}^n \to \mathbb{R}$	Weakly Convex Lipschitz Smooth
$g : \mathbb{R}^n \to \overline{\mathbb{R}}$	Weakly Convex
$L \in \mathbb{R}$	$L > 0$
$r_{\min}$	$r_{\min} \in (0, 1)$
$\rho \in \mathbb{R}$	$\rho \in (0, 1)$
$N_{\min} \in \mathbb{N}$	$N \geq 1$
$N \in \mathbb{N}$	$N \geq N_{\min}$
$\mathbb{L}_{\mathcal{S}}$	
Algorithm 0 or 0	

{alg:gmapg} 2: **Return:**

3.4 Examples of GMAPG in the literature

Example 3.4.1 (MFISTA with Armijo line search)

Algorithm 6 MFISTA with Armijo Line Search

Input: $x_{-1} \in \mathbb{R}^n, L_0 \in \mathbb{R}^n, f : \mathbb{R}^n \rightarrow \mathbb{R}, g : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$
$x_0 := y_0, t_0 := 1.$
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 3.2.1. Let's consider y_{k+1} produced the GMAPG. 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}

$$
y_{k+1} = \begin{cases} x_{k-1} + \alpha_{k+1}\alpha_k^{-1}(\tilde{x}_k - x_{k-1}) & \text{if } x_k = x_{k-1} \wedge k \ge 1, \\ \tilde{x}_k + \alpha_{k+1}(\alpha_k^{-1} - 1)(\tilde{x}_k - x_{k-1}) & \text{if } x_k = \tilde{x}_k \wedge k \ge 1, \\ \alpha_1 v_0 + (1 - \alpha_1)x_0 & \text{if } k = 0. \end{cases} \tag{3.4.1}
$$

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

{eqn:emp:result-item-2}

$$
y'_{k+1} = \begin{cases} x'_k + t_k t_{k+1}^{-1}(\tilde{x}_{k+1} - x_k) & \text{if } x'_{k+1} = x'_k \wedge k \geq 1, \\ x'_{k+1} + (t_k - 1)t_{k+1}^{-1}(\tilde{x}'_{k+1} - x'_k) & \text{if } x'_{k+1} = \tilde{x}'_{k+1} \wedge k \geq 1, \\ \tilde{x}'_1 & \text{if } k = 0. \end{cases} \tag{3.4.2}
$$

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 (3.4.1), (3.4.2) are equivalent.

Example 3.4.2 (Nesterov's monotone scheme with generic line search)
The following is (2.2.32) in Nesterov's book, phrased in our GMAPG framework.

{alg:nesterov-mono-generic-ls}

Algorithm 7 Nesterov's monotone scheme with generic line search

1: **Input:**

3.5 Practical enhancement from the Nesterov's Monotone Variant

Firstly, we will introduce some theories in for nonconvex functions. This is necessary because Nesterov's monotone variants can adapt to weakly convex objective functions. The following assumption will characterize the full scope of discussion in nonconvexity objective function.

{def:nes-monotone-scheme}

Definition 3.5.1 (nonconvex Nesterov's monotone scheme)

3.6 Restarting with function values

3.7 Applying them to large scale LP

3.8 Primal Dual Lagrangian for LP

Bibliography

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