Reading Notes

Alto

Last Compiled: May 27, 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-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:*

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

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

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

■

{thm:cnvx-pg-ineq}

Theorem 1.0.5 (convex proximal gradient inequality) *Suppose $F = f + g$ satisfies Assumption 1.0.2 such that $\mu = \mu_g \geq 0$, $\beta \geq L_f$. In addition, suppose that $f : \mathbb{R}^n \to \mathbb{R}$ has L_f Lipschitz continuous gradient, and it's $\mu_f \geq 0$ 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*

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

Proof. The Bregman Divergence of f has inequality

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

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

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

■

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. [6].

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:

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

{def:necoara-scnvx}

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

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

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

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

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

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):

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

■

Remark 2.1.8 This is Theorem 3 in Neocara et al. [6]. 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. [6], 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*

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

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

Proof. The proof is direct.

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

x^+ has descent lemma hence we have

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

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}

$$\begin{aligned} 0 &\leq f(z) - f(x^+) + \frac{L_f}{2}\|z - x\|^2 - \frac{L_f}{2}\|z - x^+\|^2 \\ &= f(z) - f(x^+) + L_f\langle z - x^+, x^+ - x\rangle + \frac{L_f}{2}\|x - x^+\|^2. \end{aligned} \tag{2.1.8}$$

Take note that when $z = x$ it has

{ineq:proj-grad2}

$$0 \leq f(x) - f(x^+) - \frac{L_f}{2}\|x - x^+\|^2. \tag{2.1.9}$$

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

$$\begin{aligned}
\|x - x_\Pi\| &\le \|x - x_\Pi^+\| \le \|x - x^+\| + \|x^+ - x_\Pi^+\| \\
&= \frac{1}{L_f}\|L_f(x - x^+)\| + \|x^+ - x_\Pi^+\| \\
\iff \|x^+ - x_\Pi^+\| &\ge \|x - x_\Pi\| - \frac{1}{L_f}\|L_f(x - x^+)\|. \qquad (2.1.10)
\end{aligned}$$

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

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:

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

Skipping some algebra, the fraction simplifies to

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

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

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

■

{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 $x \in X$:

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

■

Remark 2.1.16 It's Theorem 4 in Necoara et al. [6].

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.

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

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. ■

2.1.5 Application, KKT of linear programming

This section extends ideas in the discussion section of Necoara et al. [6].

Let X_1, X_2, Y be Hilbert spaces. Define linear mapping $E : X_1 \times X_2 \to Y := (x_1, x_2) \mapsto E_1x_1 + E_2x_2$ 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}

$$\inf_{x \in X_1 \times X_2} \{\langle -c, x\rangle \mid Ex = b, x \in \mathcal{K} \times X_2\}. \tag{2.1.12}$$

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

$$\begin{aligned} g : X_1 \times X_2 \to \mathbb{R} &:= x \mapsto \langle -c, x\rangle, \\ F : X_1 \times X_2 \to Y \times X_1 &:= (x_1, x_2) \mapsto (E_1x_1 + E_2x_2, x_1), \\ h : Y \times X_1 \to \overline{\mathbb{R}} &:= (y, z) \mapsto \delta_{\{\mathbf{0}\}}(y - b) + \delta_{\mathcal{K}}(z). \end{aligned}$$

It's not hard to identify that problem in (2.1.12) has representations

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

The dual problem of the above is given by

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

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

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

The KKT conditions give the following convex feasibility problem

$$\begin{aligned} E_1x_1 + E_2x_2 &= b, \\ E_1^*y + z &= c_1, \\ E_2^*y &= c_2, \\ \langle b, y\rangle &= \langle c_1, x_1\rangle + \langle c_2, x_2\rangle, \\ (x_1, x_2) &\in \mathcal{K} \times X_2, \\ (y, z) &\in Y \times \mathcal{K}^*. \end{aligned}$$

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^TAv 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:

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

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 [3] and Nesterov [7, 2.2.32] imposes monotonicity in function value at the iterates. Backtracking strategies from Chambolle [4] 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 discussed in Necoara et al. [6]. They showed that there exists an optimal restarting interval to achieve fast line convergence rate for all functions with quadratic growth condition. The method of restarting was improved to be parameter free while still retaining a fast linear convergence rate, see Alamo et al. [1] and Aujol et al. [2].

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

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 q_g weakly convex.

Remark 3.1.2 If a function is L smooth, it also be L weakly convex. 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:wcnvx-fxn}

Definition 3.1.3 (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 $q \geq 0$ such that the function $F + q/2\|\cdot\|^2$ is a convex function and q is the infimum of all such possible parameters.

{def:gm-for-ch2}

Remark 3.1.4 If $q = 0$, f is convex or strongly convex. If f is weakly convex, then $f + 1/2\|\cdot\|^2$ is convex and, it has dom f convex, and locally Lipschitz continuous on ri dom f.

Definition 3.1.5 (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 3.1.6 (weakly convex monotone descent)
Let $F = f + g$ satisfies Assumption 3.1.1. Let $\bar{x} = T_{\beta^{-1},f,g}(x)$. Then, for all $x \in \mathbb{R}^n$, it has the following inequality:

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

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

$$0 \leq F(x) - F(\bar{x}) - 1/\beta\|\mathcal{G}_{1/\beta}(x)\|^2.$$

The maximum amount of descent is achieved when $\beta = L + q_g$.

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

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

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

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

Optimizing $x \mapsto x - x^2(1_g - L)/2$ yields $x = (q_g - L)^{-1}$ so $\beta = q_g - L$ gives the most amount of descent. Consider any $\beta \geq (q_g - L)$:

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

■

3.1.2 smooth plus nonsmooth convex

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

Assumption 3.1.7 (convex smooth and nonsmooth) Let $F = f + g$ where $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is L Lipschitz smooth, g is convex. Suppose that $\operatorname{argmin}_{x \in \mathbb{R}^n} F(x) \neq \emptyset$.

{lemma:fitsa-pg-ineq}

Lemma 3.1.8 (proximal gradient inequality) *If $F = f + g$ satisfies Assumption 3.1.7, then for all $x \in \mathbb{R}^n, z \in \mathbb{R}^n$, define $\bar{x} = T_{L^{-1},f,g}(x)$ it has*

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

Proof. Use Theorem 1.0.5 with $\mu_f = \mu_g = 0$. ■

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 captures several monotone variants of FISTA with line search, or backtracking strategies.

In this section, we give convergence results for a unified formulation of monotone accelerated proximal gradient methods with line search or backtracking enhancements. Definition 3.2.1 gives a unified view of several combined heuristics. Theorem 3.2.4 gives a generic description for the convergence rate. Lemma 3.2.8 specializes the sequence, attains the lowest upper bound. Theorem 3.2.10 gives a concrete $\mathcal{O}(1/k^2)$ upper bound for the optimality gap and normed gradient mapping on the last iteration.

{def:generic-mapg-ls}

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

Definition 3.2.1 (GMAPG)
Initialize any $x_0, v_0 \in \mathbb{R}^n$. Let $(\alpha_k)_{k\geq 0}$ be a sequence such that $\alpha_k \in (0,1)$ $\forall k \geq 0$ and $\alpha_0 \in (0,1]$.

The algorithm makes sequences $(x_k, v_k, y_k)_{k\geq 1}$, such that for all $k = 1, 2, \ldots$ they satisfy:

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

{def:alpha-beta-rho-seq}

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

Definition 3.2.2 (alpha momentum sequence) *Let $(\alpha_k)_{k\geq 0}$ be a sequence in $\mathbb{R}$ such that $\alpha_k \in (0,1)$ for all $k \in \mathbb{N}$. Take $(L_k)_{k\geq 0}$ from Definition 3.2.1 so it has $L_k > 0$ for all $k \in \mathbb{N} \cup \{0\}$. The sequence $(\rho_k)_{k\geq 0}$ is defined as*

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

Define sequence $(\beta_k)_{k\geq 0}$ such that $\beta_0 = 1$ and for all $k \geq 1$:

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

{lemma:apg-iterates}

Lemma 3.2.3 (acceerated proximal gradient iterates relation)
The iterates $(x_k, v_k, y_k)_{k\geq 1}$ generated by Definition 3.2.1. Let $z_k = \alpha_k x^+ + (1-\alpha_k)x_{k-1}$.

Then it has for all $k \geq 1$ that:

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

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 3.2.4 (generic GMAPG convergence)
Let $F = f+g$ satisfy Assumptions 3.1.7. Take the sequence $(\alpha_k)_{k\geq 0}, (\beta_k)_{k\geq 0}$ and $(\rho_k)_{k\geq 0}$ from Definition 3.2.2. Then, for all $x^+ \in \mathbb{R}^n, k \geq 1$, the convergence rate of GMAPG (Definition 3.2.1) 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 follows, the follow facts will be used. We list them in advance, and they will be labeled during the proof.

(i) Lemma 3.2.3.

(ii) The sequence $(\alpha_k)_{k\geq 1}$ has for all $k \geq 1$, $1-\alpha_k = \alpha_k^2\alpha_{k-1}^2\rho_{k-1}$, $\alpha_k \in (0,1)$ from Definition 3.2.2.

(iii) F is convex and hence $F(z_k) \leq \alpha_k F(x^+) + (1-\alpha_k)F(x_{k-1})$ from Assumption 3.1.7.

(iv) $F(x_k) \leq F(\tilde{x}_k)$ which is true by definition of GMAPG (Definition 3.2.1).

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

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

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

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

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 1.0.5 it gives:

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

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

Remark 3.2.5 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-gm-cnvg}

Theorem 3.2.6 (generic GMAPG gradient mapping convergence) *Suppose that $F = f + g$ satisfies Assumption 3.1.7. Let the sequences (x_k, y_k, v_k) satisfy GMAPG (Definition 3.2.1), and take the momentum sequences $(\alpha_k)_{k\geq 0}, (\beta_k)_{k\geq 0}, (\rho_k)_{k\geq 0}$ from Definition 3.2.2. If in addition,*

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

(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 $k \geq 1$ the inequality:

{ineq:gmapg-generic-gm-cnvg-prt1}

$$\|\mathcal{G}_{1/L_k}(y_k)\| \leq \sqrt{\beta_k}L_kL_0\left(1 - \min(\rho_{k-1}, L_k^{-1}L_{k-1})^{1/2}\right)\|x^+ - v_0\|. \tag{3.2.1}$$

It has also:

{ineq:gmapg-generic-gm-cnvg-prt2}

$$\frac{1}{2L_0}\|\mathcal{G}_{1/L_0}(x_{-1})\|^2 \leq F(x_{-1}) - F(x_0). \tag{3.2.2}$$

Proof. (3.2.2) is direct because $x_0 = T_{1/L_0,f,g}(x_{-1})$ and $D_f(x_0, x_{-1}) \leq L_0/2\|x_0 - x_{-1}\|^2$ is assumed in the statement hypothesis, using 3.1.8 with $x = x_{-1}, z = x_{-1}$, using gradient mapping as defined in Definition 3.1.5 it has

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

We label the following results prior to their proofs for a sleeker exposition for the proof of (3.4.2).

(a) From Definition 3.2.1, the gradient mapping satisfies for all $k \geq 1$ that $\|\mathcal{G}_{1/L_k} y_k\| = L_k \alpha_k \|v_k - v_{k-1}\|$.

(b) We have $(a_k)_{k\geq 1}$ satisfying $\forall k \geq 1$ 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 $k \geq 0$ it has $\frac{\alpha_k}{\sqrt{\beta_k L_0}}\|x^+ - v_k\| \leq \|x^+ - v_0\|$.

(c) The sequence $(\alpha_k)_{k\geq 0}$ 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 $k \geq 1$.

(d) The definition of $(\beta_k)_{k\geq 0}$ from Definition 3.2.2.

Using the above intermediate results, the convergence in (3.2.1) 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 Lemma 3.1.8 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 3.2.4 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 3.2.7 The above theorem is improved from Alamo et al. [1].

{lemma:gmapg-seq-bnd}

Lemma 3.2.8 (specialized GMAPG momentum sequence) *Take sequences $(\alpha_k)_{k\ge0}, (\rho_k)_{k\ge0}, (\beta_k)_{k\ge0}$ as in Definition 3.2.2. 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 $\rho_k L_{k+1} L_k = 1$ 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

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

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 $\alpha_k \in (0,1)$, 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 $(\alpha_k)_{k\geq 0}$ has $\forall k \geq 1$ 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

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

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

Remark 3.2.9 The technique of the proof we used here is very similar to Güler [5, Lemma 2.2]. A simpler upper bound is more practical. For all $k \geq 1$ let $\widehat{L}_k = \max_{i=0,1,\dots,k} L_i$ then

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

This simplifies the convergence claim back in Theorem 3.2.4. The above inequality would work the same if we set:

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

{thm:gmapg-specialized-cnvg}

Theorem 3.2.10 (specialized GMAPG convergence rate) *Suppose that $F = f + g$ satisfy Assumption 3.1.7. Let the sequences $(x_k, v_k, v_k)_{k\geq 0}$ and $(L_k)_{k\geq 0}$ satisfy GMAPG in 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 $(\alpha_k)_{k\geq 0}$ satisfies for all $k \geq 1$: $\alpha_{k-1}^2L_k^{-1}L_{k-1}(1-\alpha_k) = \alpha_k^2$. 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)

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

(ii)

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

Proof. To see (i), use Lemma 3.2.8 and its remark to bound $(\beta_k)_{k\geq 1}$ and then, apply Theorem 3.2.4 because the assumptions of $x^+, (\alpha_k)_{k\geq 0}, (\rho_k)_{k\geq 0}$ suit. To see (ii), the convergence claim from 3.2.6 simplifies with $\widehat{L}_k \geq L_0$ and, it has

$$\|\mathcal{G}_{1/L_k}(y_k)\| \leq \left(\frac{2\sqrt{\widehat{L}_k L_0} L_k}{2+k}\right)\left(1 + \min(\rho_{k-1}, L_k^{-1} L_{k-1})^{1/2}\right)\|x^+ - v_0\| \tag{3.2.3}$$

$$= \left(\frac{2\sqrt{\widehat{L}_k L_0} L_k}{2+k}\right)\left(1 + L_k^{-1/2} L_{k-1}^{1/2}\right)\|x^+ - v_0\| \tag{3.2.4}$$

$$\leq \left(\frac{2\widehat{L}_k L_k}{2+k}\right)\left(1 + L_k^{-1/2} L_{k-1}^{1/2}\right)\|x^+ - v_0\|. \tag{3.2.5}$$

■

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.

3.3.1 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

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

{alg:armijo-ls}

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 3.1.7.

(ii) x, v are the x_k, v_k iterates in Definition 3.2.1.

(iii) α are the current α_k in Definition 3.2.1.

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

Iterates x^+, y^+ and parameters α^+, 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 $2D_f(x^+, y^+) \leq \|x^+ - y^+\|^2$ **then**
    break
  end if
  $L^+ := 2^i L^+$.
end for
Return: x^+, α^+, L^+

{alg:chambolle-btls}

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. [4]. 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.

3.3.2 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
	ρ	Number $\rho \in (0, 1)$ Number
	$G \in \mathbb{R}$	Number

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

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

{alg:nes-mono}

The above Algorithm 4 implements and adapts Nesterov's monotone scheme from Nesterov [7, 2.2.32] for GMAPG. In addition to Algorithm 3, η is a new input parameter and G has a

significance role. η is a stepsize parameter for weakly convex objective $F = f + g$ satisfying Assumption 3.1.1. G is the norm of the gradient mapping updated at at $\hat{y}$ which will be returned to the inner loop to verify exit conditions.

3.3.3 GMAPG main algorithm

Algorithm 5 GMAPG with Chambolle's backtracking

1: **Function GMAPG 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

2: $\alpha_0 := 1$.
3: $x_0, y_0, \alpha_1, L_0 := \textbf{ArmijoLS}(f, g, x_{-1}, x_{-1}, L, \alpha_0)$.
4: $\eta_0 := L_0; v_0 := x_0; G_0 = L_0(x_0 - y_0)$.
5: **for** $k := 1, 2, \ldots, N$ **do**
6: $\quad \tilde{x}_k, y_k, \alpha_{k+1}, L_k := \mathbf{L}(f, g, v_{k-1}, x_{k-1}, L_{k-1}, \alpha_k, rL_{k-1}, \rho)$.
7: $\quad \overline{L} := \max(L_k, L_{k-1})$.
8: $\quad \rho := \rho^{1/2}$ **if** $L_k > L_{k-1}$ **else** ρ.
9: $\quad G_k := L_k(\tilde{x}_k - y_k)$
10: $\quad x_k, \eta_{k+1}, G_k^+ := \mathbf{M}(f, g, \tilde{x}_k, x_{k-1}, \eta_k, G_k)$.
11: $\quad$ **if** $G_k^+ < \epsilon$ **or** ($\mathbf{E}_\chi$ **and** $k > N_{\min}$) **then**
12: $\quad\quad$ **break**
13: $\quad$ **end if**
14: **end for**
15: **Return:** $x_k, k, \overline{L}, G_k^+$

{alg:gmapg}

The above Algorithm 5 is an implementation of GMAPG in Definition 5. Parameters r, ρ are chosen in the discretion of the practitioners. For example, we chose $r = 0.4, \rho = 2^{1/1024}$.

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

{alg:mfista-armijo}

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 $k \geq 0$

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

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)
Suppose $F = f + g$ satisfies Assumption 3.1.1. Let $L_0 \geq L$. Let $(\alpha_k)_{k\geq 0}$ with $\alpha_0 = 1$ and, it satisfies for all $k \geq 1$: $L_k^{-1} L_{k-1}\alpha_{k-1}^2(1-\alpha_k) = \alpha_k^2$. 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 $F(x_0) \leq F(x_{-1})$. The algorithm is defined by sequences $(y_k, v_k, x_k)_{k\geq 1}$ and $(\tilde{x}_k, \hat{y}_k)_{k\geq 1}$ such that they all satisfy:

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

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 3.5.2 (convergence of Nesterov's monotone scheme nonconvex) *Suppose that the sequences $(y_{k+1}, v_k, x_k)_{k\geq 0}$ and $(\hat{y}_k, \tilde{x}_k)_{k\geq 0}$, $(\alpha_k)_{k\geq 0}$ satisfy Definition 4. Assume that F is bounded below with $F^+ := \inf_x F(x)$. Then for all $N \geq 1$ it has*

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

Here, $\overline{\eta}_k = \max_{i=0,\dots,k} \eta_i$. 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 3.1.6 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 3.1.6 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 η_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 3.5.3 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 3.1.7.

3.6 Restarting with function values for linear convergence

In this section, we show that restarting GMAPG with a well suited conditions yields fast linear global convergence. We adapt and improve prior theories from Alamo [1] for our GMAPG 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 3.6.1 (quadratic growth condition) Let $F = f + g$ satisfies Assumption 3.1.7 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

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

The following proposition about line search will furnish convergence results from previous section.

{prop:bnded-lip-ls}

Proposition 3.6.2 (Lipschitz line search estimates are bounded) *Suppose that $F = f+g$ satisfies Assumption 3.1.7. Choose such F for Algorithm 5 so, it generates the sequence $(L_k)_{k\geq 0}$. Then, the sequence $(\hat{L}_k)_{k\geq 1}$ from Theorem 3.2.10 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. A line search is triggered in Algorithm 1, 2 if and only if $L_{k+1} = 2L_k$ for some $k \geq 0$ and, there exists some $x \in \mathbb{R}^n, y \in \mathbb{R}^n$ such that $D_f(x, y) > L_k\|x - y\|^2$. 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 $x, y \in \mathbb{R}^n$. Hence, $L_{k+1} \leq L_k$ because a line sarch is never triggered.

For contradiction let's assume that there exists $k \geq 1$ such that a line search is triggered for L_k and, $L_{k+1} = 2L_k > 2L$, then $L_k > L$, but we just showed that this 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 $2L$. ■

Let's introduce our first set of restart conditions which denote it by $\mathbf{E}_\chi^a$. $\mathbf{E}_\chi^a$ only uses function values on iterates available to the inner loop to determine the exit conditions, and it achieves global fast linear convergence. The inner loop algorithm is described in Algorithm 5, let k denote the inner loop counter and define $m = \lfloor k/2 \rfloor + 1$ the conditions are:

{ineq:rgmapg-exit-cond}

$$\mathbf{E}_\chi^a \iff f(x_m) - f(x_k) \leq \exp(-1)(f(x_0) - f(x_m)). \tag{3.6.1}$$

Algorithm 8 Linear convergence restarted GMAPG

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
$L := 1$	$L > 0$
$r := 0.5$	$r \in (0, 1)$
$\rho := 2^{1/1024}$	$\rho \in (0, 1)$

$n_0 := 0; z_0 = x_{-1}$.
$z_1, p_0, \overline{L}_1 := \mathbf{GMAPG}(f, g, x_{-1}, L, r_j, \rho, N_{\min} = n_0, N = M, \epsilon, \mathbf{L}, \mathbf{M}, \mathbf{E}_\chi^a)$.
$n_1 := p_0$.
$M := M - n_1$
for $j = 1, 2, \ldots, M$ **do**
  $z_{j+1}, p_j, \overline{L}_{j+1}, G_k^+ := \mathbf{GMAPG}(f, g, z_j, \overline{L}_j, r, \rho, N_{\min} = n_j, N = M, \epsilon, \mathbf{L}, \mathbf{M}, \mathbf{E}_\chi^a)$.
  $M := M - p_j$.
  if $M \leq 0$ **or** $G_k^+ \leq \epsilon$ **then**
    break
  end if
  $\overline{L}_{j+1} = \max(\overline{L}_j, \overline{L}_{j+1})$.
  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
end for

{alg:linear-rgmapg}

Algorithm 8 implements a restarted GMAPG with condition $\mathbf{E}_\chi^a$ and, it has a fast linear convergence rate. The following observation on it is crucial to take notice for the convergence rate.

n_j is used as the lower bound $N_{\min}$ for the GMAPG inner loop described in Algorithm 5 at iteration j and, p_j is the actual iteration underwent by the inner loop and, it has $p_j \geq n_j$. n_{j+1} is the minimum iteration for the next execution of GMAGP, and it has $n_{j+1} = p_j$ if the "if statement" isn't triggered otherwise $n_{j+1} = 2p_j$. Consequently, it has either $p_{j+1} \geq n_{j+1} = 2p_j$ or $p_{j+1} \geq n_{j+1} = p_j$.

We now introduce the following notations for the convergence proof. Since the restarted GMAPG consist of executions of an outer loop in Algorithm 8 in variable j and, an inner loop in Algorithm 5 in variable k, we denote $z_{j,k}$ for the iterates x_k in the inner loop and z_j in the outer loop together. We make the choice to have $z_{j+1,-1} = z_{j,p_j} = z_{j+1}$ for consistencies across theorems from previous sections. For example in line 7 of Algorithm 8 at the j iteration, the inner loop returns x_{p_j} as the last iterate. Denote x_{p_j} as z_{j,p_j} and, it's assigned to z_{j+1} by the outer loop so $x_{p_j} = z_{j,p_j} = z_{j+1}$. It continues and z_{j+1} will be the initial iterate pass into the inner loop for the $j+1$ iteration, and hence $z_{j+1} = z_{j+1,-1}$.

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

{lemma:prog-ratio}

Lemma 3.6.3 (maximum iteration needed for an optimality gap ratio) *Suppose that $F = f + g$ satisfies Assumption 3.1.7 so $F^+ := \inf_x F(x)$ and, X^+ is the set of minimizers. Let the sequence $(x_k)_{k\geq -1}$ be generated by GMAPG (Definition 5). If in addition, F satisfies the quadratic growth condition in Assumption 3.6.1, then*

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

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 3.2.10 it has

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

The second inequality comes from Assumption 3.6.1 directly. Suppose that $k \geq 2\sqrt{1+e}\left(\mu\widehat{L}_k^{-1}\right)^{-1/2}$. 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:

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

Continuing it has

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

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

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

■

{lemma:rgmapg-inner-bnds}

Continuing with the quadratic growth assumption, the following lemma states the fact that p_j in Algorithm 8 is bounded above and there exists a for p_j such that $n_{j+1+k} = p_{j+k}$ for all $k \geq 0$ until it terminates.

Lemma 3.6.4 (inner iteration is bounded above) *Suppose that $F = f + g$ satisfies Assumption 3.1.7 so $F^+ := \inf_x F(x)$ and, X^+ is the set of minimizers. Consider any $j \geq 1$ iteration experienced by the outer loop. Let $(z_{j,k})_{k\geq 0}^{p_j}$ be the sequence generated by Algorithm 8 in the j th iteration of outer loop. Define $\bar{p} := \frac{4\sqrt{2L(1+e)}}{\sqrt{\mu}}$, then $p_j \leq \bar{p}$.*

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: $p_0 \leq p_1 \leq \bar{p}$. 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 $j \geq 0$, $p_j \leq \bar{p}$.

Proof of (i). Recall exit condition in (3.6.1) has $m = \lfloor k/2 \rfloor + 1$. Starting with the statement hypothesis it has $k \geq barp$ 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 3.6.2 which has $\widehat{L}_k \leq 2L$. Observe that the inequality allow us to apply Lemma 3.6.3 with $m = k$ which yields:

$$e^{-1} \geq \frac{F(z_{j,m}) - F^+}{F(z_{j,-1}) - F(z_{j,m})} \geq \frac{F(z_{j,m}) - F(z_{j,k})}{F(z_{j,-1}) - F(z_{j,m})} \implies \mathbf{E}_\chi^a.$$

Observe that line 11 of Algorithm 5 exits as soon as possible if E_χ^a is true and $k > N_{\min} = n_j$ and, $G_k \leq \epsilon$ will only cause it to exit earlier therefore $p_j \leq \max(\bar{p}, n_j)$.

Proof of (ii). Inductively assume that $p_{j-1} \leq p_j \leq \bar{p}$. If $p_{j-1} \leq \bar{p}/2$ then Line 14, 16 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 3.6.2 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 3.6.3 which yields

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

Therefore, there is no doubling at line 14 of Algorithm 8, hence $n_{j+1} \leq \bar{p}$ still. ■

We just show that the sequence n_j is bounded above, and it must have a limit because it's also monotone increasing, 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 that happens, the algorithm will always terminate after a finite number of iterations of the outer loop.

Lemma 3.6.5 (bounds on outer iteration counts) ...

Theorem 3.6.6 (bounds on the total iterations) ...

3.7 Applying them to large scale LP

3.8 Hoffman Bounds and infeasibility detection

Chapter 4

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

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

[6] 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.

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