First Order Nonsmooth Optimization: Algorithm Design, Variational analysis, and Applications

Hongda Li

Department of Mathematics
University of British Columbia,
Okanagan Campus.

January 21, 2025

Contents

1 Introduction

Let $\mathbb{R}^n$ be the ambient space. We consider

$$\min_{x\in\mathbb{R}^n} \{F(x) := f(x) + g(x)\}. \tag{1.1}$$

Unless specified, assume $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipscthiz smooth $\mu \geq 0$ strongly convex and $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex. This type of problem is referred to as additive composite problems in the literature.

Our ongoing research concerns accelerated proximal gradient type method for solving (1). In the expository writing by Walkington [21], a variant for of accelerated gradient method for strongly convex function f is discussed. We had some lingering questions after reading it.

(i) Do there exist a unified description for the convergence for both variants of the algorithms?

(ii) Is it possible to attain fast convergence rate without knowledge about the strong convexity of function f?

(iii) Is it possible to describe the convergence of function value for momentum sequences that are much weaker than the Nesterov's rule?

The good news is we have definitive answers for all questions in our complete draft paper.

In this proposal we explore the Goldilocks zones between these topics: Nesterov's acceleration, algorithms with inexact evaluation of the proximal point operator, and applications. We want to identify the "chemistry" between various properties of functions and the affect it has on the designs/behaviors of first order algorithm for continuous optimization.

Organizations now follows. Section 3, proposes the method of "Relaxed Weak Accelerated Proximal Gradient (R-WAPG)" as the foundation to describe several Euclidean variants of Accelerated Proximal Gradient (APG) method in the literatures. The convergence analysis of R-WAPG gives convergence of APG the momentum is much weaker compare to the Nesterov's update rule. R-WAPG can is generic and can describe well-established variants of FISTA in the literatures with alternative choices of step sizes and beyond.

Section 5 proposes a practical algorithm that exploits a specific term in the proof of R-WAPG to achieve faster convergence for solving (1) without restarting and knowing parameter L, μ in prior. Numerical experiments are presented which shows the competitiveness of our algorithm. Potential future direction of research is given by the end of the section for the R-WAPG framework.

Section 6 complements report completed for MATH 590 Fall Winter 2024. It's based on a series of recent papers [14, 15, 22] about Catalyst Meta Acceleration for First Order Variance Reduced Methods. The end of Section 6 points out potential future direction of research of Catalyst Acceleration.

2 Preliminaries

This section contains the basics of contents from convex optimization, and variational analysis. Throughout, we adopt the notation $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty, -\infty\}$.

2.1 Fundamentals in non-convex analysis

Let the ambient space be $\mathbb{R}^n$ equipped with inner product $\langle \cdot, \cdot \rangle$ and 2-norm $\|\cdot\|$. Let O be an open subset of $\mathbb{R}^n$, the weakest assumption we make for objective function $F : O \subseteq \mathbb{R}^n \to \mathbb{R}$ of an optimization problem is local Lipschitz continuity. The assumption of local Lipschitz continuity is weak enough to describe most problems in applications, and strong enough to avoid most pathologies in analysis.

Definition 2.1 (Local Lipschitz continuity) *Let $F : O \subseteq \mathbb{R}^n \to \mathbb{R}$ be Locally Lipschitz and O is an open set. Then for all $\bar{x} \in O$, there exists a Neighborhood: $\mathcal{N}(\bar{x})$ and $K \in \mathbb{R}$ such that for all $x, y \in \mathcal{N}(\bar{x})$: $|F(x) - F(y)| \leq K\|x - y\|$.*

Definition 2.2 (Regular subgradient) *Let $F : O \subseteq \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz and $\bar{x} \in O$. The regular subdifferential at $\bar{x}$ is defined as*

$$\widehat{\partial} F(\bar{x}) := \left\{ v \in \mathbb{R}^n \;\middle|\; \liminf_{\bar{x} \neq x \to \bar{x}} \frac{F(x) - F(\bar{x}) - \langle v, x - \bar{x} \rangle}{\|x - \bar{x}\|} \geq 0 \right\}.$$

Remark 2.3 Definition taken from Definition 4.3.1 from Pang, Cui's book [29]

Definition 2.4 (Limiting subgradient) *Let $F : O \subseteq \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz and $\bar{x} \in O$. The limiting subdifferential at $\bar{x}$ is defined as*

$$\partial F(\bar{x}) := \left\{ v \in \mathbb{R}^n \;\middle|\; \exists x_k \to \bar{x}, v_k \to v : v_k \in \widehat{\partial} F(x_k) \; \forall k \in \mathbb{N} \right\}.$$

Remark 2.5 Definition taken from Definition 4.3.1 from Pang, Cui's book [29]

Definition 2.6 (Weakly convex function) *$F : \mathbb{R}^n \to \overline{\mathbb{R}}$ is μ weakly convex if and only if $F + \frac{\mu}{2}\|\cdot\|^2$ is convex.*

Definition 2.7 (Bregman divergence) *Let $F : O \subseteq \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. Then the Bregman divergence of F is defined as:*

$$D_F(x, y) : O \times \operatorname{dom}(\partial F) \to \mathbb{R} := F(x) - F(y) - \langle \nabla F(y), x - y \rangle.$$

2.2 Fundamentals in convex analysis

This section introduces the classics and basics of convex analysis. Define F to be closed, proper and convex in this section. When F is convex, the limiting subgradient and the regular subgradient reduced to the following:

$$\partial F(x) = \{v \in \mathbb{R}^n \mid \forall y \in \mathbb{R}^n \ : F(y) - F(x) \geq \langle v, y - x \rangle\}.$$

A convex function is locally Lipschitz in the relative interior of its domain, denoted as $\operatorname{ri}(\operatorname{dom}(F))$. So it has $\operatorname{ri}(\operatorname{dom} F) \subseteq \operatorname{dom}(\partial F) \subseteq \operatorname{dom} F$.

- ☐ Add Rockafellar's red book
- ☐ Cite Section 10 (pg 82) of that book

When we say $F : \mathbb{R}^n \to \mathbb{R}$ is L Lipschitz smooth function, it means that there exists L such that for all $x \in \mathbb{R}^n, y \in \mathbb{R}^n$, it has:

$$\|\nabla F(x) - \nabla F(y)\| \leq L\|x - y\|.$$

When F convex, then it has descent lemma:

$$(\forall x \in \mathbb{R}^n)(\forall y \in \mathbb{R}^n) : 0 \leq F(x) - F(y) - \langle \nabla f(y), x - y \rangle \leq \frac{L}{2}\|x - y\|^2.$$

The converse holds under convexity as well. The definitions that follow narrow things further for future discussions.

Definition 2.8 (Strong convexity) *A function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ is $\mu \geq 0$ strongly convex if and only if for all $y \in \operatorname{dom}(\partial F)$, $x \in \mathbb{R}^n$:*

$$(\forall v \in \partial F(x)) \quad F(x) - F(y) \geq \langle v, x - y \rangle + \frac{\mu}{2}\|x - y\|^2.$$

Lemma 2.9 (Quadratic growth from strong convexity) *If F is $\mu \geq 0$ strongly convex, $\bar{x}$ is a minimizer of F. Then for all $x \in \mathbb{R}^n$*

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

Remark 2.10 The minimizer is unique whenever $\mu > 0$. For contradiction, assume $x \neq \bar{x}$ is another minimizer, then $F(x) \neq F(\bar{x})$, which is a direct contradiction. This condition is called quadratic growth over the set of minimizer, it is much weaker than strong convexity.

2.2.1 Smooth, nonsmooth additive composite

This section zooms in further into the case of additive composite objective $F := f + g$. Assume f is L Lipschitz smooth and $\mu \geq 0$ strongly convex, $g : \mathbb{R} \to \overline{\mathbb{R}}$ is closed convex. For all $\beta \geq 0$, define the proximal gradient, proximal point model functions as a mapping of $\mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$:

$$\widetilde{\mathcal{M}}_F^{\beta^{-1}}(x; y) := g(x) + f(y) + \langle \nabla f(y), x - y \rangle + \frac{\beta}{2}\|x - y\|^2,$$
$$\mathcal{M}_F^{\beta^{-1}}(x; y) := F(x) + \frac{\beta}{2}\|x - y\|^2.$$

Under the assumptions of this section, $\widetilde{\mathcal{M}}_F^{\beta^{-1}}(\cdot; y), \mathcal{M}_F^{\beta^{-1}}(\cdot; y)$ are both $\beta+\mu$ strongly convex.

Definition 2.11 (Proximal gradient operator) *Define the proximal gradient operator T_L on all $y \in \mathbb{R}^n$:*

$$T_L y := \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ g(x) + f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|^2 \right\}.$$

Remark 2.12 Under the assumption of this section, the mapping T_L is a single-valued mapping on $\mathbb{R}^n$ and it's a $3/2$ averaged operator.

Definition 2.13 (Gradient mapping operator) *Define the gradient mapping operator $\mathcal{G}_L$ on all $y \in \mathbb{R}^n$:*

$$\mathcal{G}_L(y) := L(y - T_L y).$$

Lemma 2.14 (Proximal gradient model function)
For all $x \in \mathbb{R}^n$, $\beta > 0$, it has:

$$\widetilde{\mathcal{M}}_F^{\beta^{-1}}(x; y) = \mathcal{M}_F^{\beta^{-1}}(x; y) - D_f(x, y).$$

Lemma 2.15 (A favorable property of gradient mapping) *For any $x \in \mathbb{R}^n$. Then there exists $v \in \partial g(T_L x)$ such that $\mathcal{G}_L(x) = v + \nabla f(x)$.*

Lemma 2.16 (The proximal gradient inequality) *For all $y \in \mathbb{R}^n$, $x \in \mathbb{R}^n$, it has:*

$$(\forall x \in \mathbb{R}^n) \quad F(x) - F(T_L y) - \langle L(y - T_L y), x - y \rangle - \frac{\mu}{2}\|x - y\|^2 - \frac{L}{2}\|y - T_L y\|^2 \geq 0.$$

Remark 2.17 This lemma is proved in our draft paper.

2.3 Nesterov's estimating sequence technique

The derivation of APG was originally conceived by Nesterov's estimating sequence technique. We emphasize that the technique derives the algorithm and proves its convergence rate.

The method is widespread in the literatures, and the ideas behind it are tremendously useful. Güler [12] used it to derive an accelerated proximal point method, which was instrumental to develop the Catalyst Acceleration framework. Nesterov [17] used it to conceive the accelerated cubic regularized Newton's method. In (6.1.19) of Nesterov's book [20], it's used to derive a method of accelerated mirror descent. Finally, in Geovani N. et al [11], they used it to derive an accelerated Newton's method for convex composite objective function.

The definition of the estimating sequence that follows is based on our own understanding of the estimating sequence.

Definition 2.18 (Nesterov's estimating sequence) *For all $k \geq 0$, let $\phi_k : \mathbb{R}^n \to \mathbb{R}$ be a sequence of functions. We call this sequence of functions a Nesterov's estimating sequence when it satisfies conditions:*

(i) *There exists another sequence $(x_k)_{k\geq 0}$ such that for all $k \geq 0$ it has $F(x_k) \leq \phi_k^* := \min_x \phi_k(x)$.*

(ii) *There exists a sequence of $(\alpha_k)_{k\geq 0}$ where $\alpha_k \in (0,1)\ \forall k \geq 0$ such that for all $x \in \mathbb{R}^n$ it has $\phi_{k+1}(x) - \phi_k(x) \leq -\alpha_k(\phi_k(x) - F(x))$.*

Observation 2.19 *In general, identifying the sequence $(x_k)_{k\geq 0}$ is non-trivial. But in case it can be found, the method of estimating sequence gives us the convergence rate described by the sequence $(\alpha_k)_{k\geq 0}$, and a candidate algorithm that generates the sequence $(x_k)_{k\geq 0}$. It's two birds with one stone.*

If we define ϕ_k, $\Delta_k(x) := \phi_k(x) - F(x)$ for all $x \in \mathbb{R}^n$ and assume that F has minimizer x^*. Then observe that $\forall k \geq 0$:

$$\begin{aligned}
(\forall x \in \mathbb{R}^n)\ \Delta_k(x) &= \phi_k(x) - F(x) \geq \phi_k^* - F(x), \\
x = x_k \implies \Delta_k(x_k) &\geq \phi_k^* - F(x_k) \geq 0; \\
x = x_* \implies \Delta_k(x_*) &\geq \phi_k^* - F_* \geq F(x_k) - F_* \geq 0.
\end{aligned}$$

The function $\Delta_k(x)$ is non-negative at points: x_*, x_k. We can derive the convergence rate of $\Delta_k(x^*)$ because (ii) says $\forall x \in \mathbb{R}^n$:

$$\begin{aligned}
\phi_{k+1}(x) - \phi_k(x) = \phi_{k+1}(x) - F(x) - (\phi_k(x) - F(x)) &\leq -\alpha_k(\phi_k(x) - F(x)) \\
\iff \Delta_{k+1}(x) - \Delta_k(x) &\leq -\alpha_k \Delta_k(x) \\
\iff \Delta_{k+1}(x) &\leq (1-\alpha_k)\Delta_k(x).
\end{aligned}$$

Unrolling the above recursion it yields:

$$\Delta_{k+1}(x) \leq (1-\alpha_k)\Delta_k(x) \leq \cdots \leq \left(\prod_{i=0}^{k}(1-\alpha_i)\right)\Delta_0(x).$$

Finally, by setting $x = x^*$, $\Delta_k(x^*)$ is non-negative and using (i) it gives:

$$F(x_k) - F(x^*) \leq \phi_k^* - F(x^*) \leq \Delta_k(x^*) = \phi_k(x^*) - F(x^*) \leq \left(\prod_{i=0}^{k}(1-\alpha_i)\right)\Delta_0(x^*).$$

3 Unifying NAG, and weakening the sequence assumption for convergences (WORK IN PROGRESS)

This section is based on our unpublished draft paper.

3.1 Contributions

Inspired specifically by the technique of Nestrov's estimating sequence [20], firstly we present a unified framework of Accelerated Proximal Gradient (APG) which we call Relaxed Weak Accelerated Proximal Gradient (R-WAPG) in Section 3.3. It has the ability to upper bound $F(x_k) - F(x^*)$ for sequences $(t_k)_{k\geq 0}$ that follows a rule much weaker than Nesterov's update rule. In addition to a convergence claim of $F(x_k) - F(x^*)$ for a much more flexible choice of $(t_k)_{k\geq 1}$. Secondly, we present an alternative to restarting that performs well empirically inspired by a small detail in the convergence proof of R-WAPG. It also has descriptive power to describe several variants of FISTA in the literatures.

Our contributions are two folds, theoretical and practical. Our results are based the assumption $F = f + g$ where $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex proper and closed, and f is an L-Lipschitz smooth and $\mu \geq 0$ strongly convex function.

A summary of our main results follow. Nesterov's acceleration extrapolate $y_{k+1} = x_{k+1} + \theta_{k+1}(x_{k+1} - x_k)$ where $\theta_{k+1} = (t_k - 1)/t_{k+1} \in (0, 1)$ is the "momentum". The choices for θ_k varies for different variants of the accelerated proximal gradient algorithm. In Chambolle, Dossal [5], it has $t_k = (n + a - 1)/a$ for all $a > 2$ which gives weak convergence of the iterates x_k in Hilbert space. In Chapter 10 of Beck's Book [2], a variant called V-FISTA can achieve the faster linear convergence rate: $\mathcal{O}((1 - \sqrt{\mu/L})^k)$ on the optimality gap for $\mu > 0$ strongly convex F. V-FISTA has $\theta_t = (\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)$ where $\kappa = \mu/L$.

We relax the traditional choice of the sequence θ_k in Equation **??** and showed an upper bound of the optimal gap. Let $(\alpha_k)_{k\geq 0}, (\rho_k)_{k\geq 0}$ be two sequences that satisfy

$$\begin{aligned}
&\alpha_0 \in (0, 1], \\
&\alpha_k \in (\mu/L, 1) \quad (\forall k \geq 1), \\
&\rho_k := \frac{\alpha_{k+1}^2 - (\mu/L)\alpha_{k+1}}{(1-\alpha_{k+1})\alpha_k^2} \quad \forall (k \geq 0).
\end{aligned}$$

Our first main result shows that if $\theta_{k+1} = (\rho_k \alpha_k (1-\alpha_k)/(\rho_k \alpha_k^2 + \alpha_{k+1}))$, using the R-WAPG we proposed in Definition 3.5 with Proposition 3.6, **??**, we can show that the gap $F(x_k) - F(x^*)$ is bounded by:

$$\mathcal{O}\left(\left(\prod_{i=0}^{k-1} \max(1, \rho_k)\right) \prod_{i=1}^{k} (1-\alpha_i)\right).$$

Our second main result shows that for any $\alpha_k \geq 0$ the choice of sequence $\alpha_k = a/(a+k)$ results in $\rho_k > 1$ for all $k \in \mathbb{N}$ such that R-WAPG reduces to a variant of FISTA proposed in Chambolle, Dossal [5], and we are able to show the same convergence rate in Theorem 4.4. When $\rho_k = 1, \mu = 0$, R-WAPG reduces perfectly to FISTA by Beck [2], if $\mu > 0, \rho_k = 1$, it reduces to the V-FISTA by Beck [2]. In Theorem 4.6, it demonstrates that R-WAPG frameworks gives a linear convergence claim for all fixed momentum method where $\alpha_k := \alpha \in (\mu/L, 1)$ and F is $\mu > 0$ strongly convex. Finally, we did the tedious work and present three equivalent forms of R-WAPG in Section 3.4 that are comparable to notable Euclidean variants of FISTA and beyond. This shows the descriptive power of our R-WAPG framework.

Our practical contribution is an algorithm inspired by a detail in our convergence proof which we call it "Parameter Free R-WAPG" (See Algorithm 1). The algorithm is parameter free, meaning that it doesn't require knowing L, μ in advance, and it determines the value of θ_t by estimating the local concavity using iterates y_k, y_{k+1} from the Bregman divergence of f with minimal computational cost. We conducted ample amount of numerical experiments to show that it has a favorable convergence rate in practice and behaves similarly to the FISTA with monotone restart.

Section **??** introduces the proximal gradient inequality which is instrumental to developing the stepwise convergence claim in the next section. Section 3.2 states and prove an inequality based on a generic iterative algorithm for just one iteration. Section 3.3 formulates the full R-WAPG algorithm and characterize sufficient conditions for R-WAPG sequence to derive the optimality upper bound in Proposition 3.6. Section 3.4 gives three equivalent representations of the R-WAPG algorithms that are comparable to instances of APG found in the literatures. Section 4 formulates FISTA, and V-FISTA sequences as instance of the R-WAPG sequences. The section proves the convergence rate of several variants of FISTA using the equivalent forms introduced in Section 3.4 and the convergence rate developed Section 3.3. Finally, in

Section 5 gives a formulation of a parameter free version of R-WAPG algorithm and showcase the numerical experiments for regression, LASSO. The numerical experiments expose very interesting behaviors of the algorithm of our own creativity.

3.2 Stepwise formulation of weak accelerated proximal gradient

The goal of this section is to build the R-WAPG algorithm which is described in the Definition 3.5 of the next section.

Definition 3.2 which describes what happens at a single iteration of the R-WAPG algorithm. It defines a procedure of generating x_{k+1}, v_{k+1} given any x_k, v_k. Proposition 3.3 states the inequality that describes a decreasing quantity that involves $F(x_k), F(x_{k+1})$ at each single iteration.

Assumption 3.1 Given x_k, y_k, v_k where $k \in \mathbb{Z}_+$, we define the following quantities

$$g_k := L(y_k - T_L y_k), \tag{3.1}$$

$$l_F(x; y_k) := F(T_L y_k) + \langle g_k, x - y_k \rangle + \frac{1}{2L}\|g_k\|^2, \tag{3.2}$$

$$\epsilon_k := F(x_k) - l_F(x_k; y_k), \tag{3.3}$$

Observe that by convexity of F, $\epsilon_k \geq 0$ for all $x_k, L > 0$. To see, use Theorem 2.16 and let $y = y_k, x = x_k$ which gives:

$$F(x_k) - F(T_L y_k) - \langle L(y_k - T_L y_k), x_k - y_k \rangle - \frac{L}{2}\|y_k - T_L y_k\|^2 - \frac{\mu}{2}\|x_k - y_k\|^2 \geq 0$$
$$\iff F(x_k) - F(T_L y_k) - \langle g_k, x_k - y_k \rangle - \frac{1}{2L}\|g_k\|^2 \geq 0.$$

The proposition follows provides upper bound to $F(x_{k+1})$ in relations to $F(x_k)$.

Definition 3.2 (Stepwise weak accelerated proximal gradient)
Assume $0 \leq \mu < L$. Fix any $k \in \mathbb{Z}_+$. For any $(v_k, x_k), \alpha_k \in (0,1), \gamma_k > 0$, let $\hat{\gamma}_{k+1}$, and vectors y_k, v_{k+1}, x_{k+1} be given by:

$$\hat{\gamma}_{k+1} = (1 - \alpha_k)\gamma_k + \mu\alpha_k, \tag{3.4}$$

$$y_k = (\gamma_k + \alpha_k\mu)^{-1}(\alpha_k\gamma_k v_k + \hat{\gamma}_{k+1}x_k), \tag{3.5}$$

$$g_k = \mathcal{G}_L y_k, \tag{3.6}$$

$$v_{k+1} = \hat{\gamma}_{k+1}^{-1}(\gamma_k(1 - \alpha_k)v_k - \alpha_k g_k + \mu\alpha_k y_k), \tag{3.7}$$

$$x_{k+1} = T_L y_k. \tag{3.8}$$

Proposition 3.3 (Stepwise Lyapunov)
Let $k \in \mathbb{Z}_+$, $R_k \in \mathbb{R}$. Given any v_k, x_k and $\gamma_k > 0$ and $v_{k+1}, x_{k+1}, y_k, \hat{\gamma}_{k+1}, \alpha_k$ that satisfies Definition 3.2. Define:

$$R_{k+1} := \frac{1}{2}\left(L^{-1} - \frac{\alpha_k^2}{\hat{\gamma}_{k+1}}\right)\|g_k\|^2 + (1-\alpha_k)\left(\epsilon_k + R_k + \frac{\mu\alpha_k\gamma_k}{2\hat{\gamma}_{k+1}}\|v_k - y_k\|^2\right). \tag{3.9}$$

Then for all $x^* \in \mathbb{R}^n$, we have:

$$F(x_{k+1}) - F(x^*) + R_{k+1} + \frac{\hat{\gamma}_{k+1}}{2}\|v_{k+1} - x^*\|^2 \le (1-\alpha_k)\left(F(x_k) - F(x^*) + R_k + \frac{\gamma_k}{2}\|v_k - x^*\|^2\right). \tag{3.10}$$

3.3 R-WAPG and its convergence rates

In this section we propose Relaxed Weak Accelerated Proximal Gradient (R-WAPG), see Definition 3.5. R-WAPG algorithm generates iterates (x_k, y_k, v_k) and admits an upper bound on $F(x_k) - F^*$ described in Proposition 3.3. Definition 3.4 introduces the concept of an R-WAPG sequences: $(\alpha_k)_{k\ge0}, (\rho_k)_{k\ge0}$ which is crucial. The sequences parameterize the R-WAPG algorithm stated in Definition 3.5, it connects the step-wise formulation of R-WAPG (Definition 3.2) and can describe the convergence claim of R-WAPG in Proposition 3.6. In the next section, it continues to play a crucial role in describing several equivalent forms of the R-WAPG algorithm, and their corresponding convergence claim.

Definition 3.4 (R-WAPG sequences)
Assume $0 \le \mu < L$. The sequences $(\alpha_k)_{k\ge0}, (\rho_k)_{k\ge0}$ are valid for R-WAPG if all the following holds:

$$\begin{aligned}
&\alpha_0 \in (0, 1],\\
&\alpha_k \in (\mu/L, 1) \quad (\forall k \ge 1),\\
&\rho_k := \frac{\alpha_{k+1}^2 - (\mu/L)\alpha_{k+1}}{(1-\alpha_{k+1})\alpha_k^2} \quad \forall (k \ge 0).
\end{aligned}$$

*We call $(\alpha_k)_{k\ge0}, (\rho_k)_{k\ge0}$ the **R-WAPG Sequences**.*

We now give Relaxed Weak Accelerated Proximal Gradient in details.

Definition 3.5 (Relaxed weak accelerated proximal gradient (R-WAPG))
Choose any $x_1 \in \mathbb{R}^n, v_1 \in \mathbb{R}^n$. Let $(\alpha_k)_{k\ge0}, (\rho_k)_{k\ge0}$ be given by Definition 3.4. The algorithm generates a sequence of vector $(y_k, x_{k+1}, v_{k+1})_{k\ge1}$ for $k \ge 1$ by the procedures:

For $k = 1, 2, 3, \cdots$

$$
\begin{aligned}
\gamma_k &:= \rho_{k-1} L \alpha_{k-1}^2, \\
\hat{\gamma}_{k+1} &:= (1-\alpha_k)\gamma_k + \mu\alpha_k = L\alpha_k^2, \\
y_k &= (\gamma_k + \alpha_k \mu)^{-1}(\alpha_k \gamma_k v_k + \hat{\gamma}_{k+1} x_k), \\
g_k &= \mathcal{G}_L y_k, \\
v_{k+1} &= \hat{\gamma}_{k+1}^{-1}(\gamma_k(1-\alpha_k)v_k - \alpha_k g_k + \mu\alpha_k y_k), \\
x_{k+1} &= T_L y_k.
\end{aligned}
$$

Here is the main result of this section.

Proposition 3.6 (R-WAPG convergence claim)
Fix any arbitrary $x^* \in \mathbb{R}^n, N \in \mathbb{N}$. *Let vector sequence* $(y_k, v_k, x_k)_{k\geq 1}$ *and R-WAPG sequences* α_k, ρ_k *be given by Definition 3.5. Define* $R_1 = 0$ *and suppose that for* $k = 1, 2, \ldots, N$, *we have* R_k *recursively given by:*

$$
R_{k+1} := \frac{1}{2}\left(L^{-1} - \frac{\alpha_k^2}{\hat{\gamma}_{k+1}}\right)\|g_k\|^2 + (1-\alpha_k)\left(\epsilon_k + R_k + \frac{\mu\alpha_k\gamma_k}{2\hat{\gamma}_{k+1}}\|v_k - y_k\|^2\right).
$$

Then for all $k = 1, 2, \ldots, N$:

$$
\begin{aligned}
&F(x_{k+1}) - F(x^*) + \frac{L\alpha_k^2}{2}\|v_{k+1} - x^*\|^2 \\
&\leq \left(\prod_{i=0}^{k-1} \max(1, \rho_i)\right)\left(\prod_{i=1}^{k}(1-\alpha_i)\right)\left(F(x_1) - F(x^*) + \frac{L\alpha_0^2}{2}\|v_1 - x^*\|^2\right).
\end{aligned}
$$

3.4 Equivalent representations of R-WAPG

This section reduces Definition 3.5 into simpler forms that are comparable to what commonly appears in the literatures. In the literatures, variants of Accelerated Proximal Gradient algorithm such as FISTA, V-FISTA has different representations. This shows that R-WAPG provides a unified framework. These equivalent representations are listed in Definition 3.7, 3.9 and 3.11. These forms are equivalent under a subset of initial conditions.

They are comparable to existing APG algorithms in the literatures such as Excersise 12.1 in Ryu, Yin [24], Similar Triangle from from Lee et al. [13], Ahn Sra [1] and momentum form of (2.2.19) in Nesterov [20]. Specific instances of Accelerated Proximal Gradient algorithm that has the same form as the Definition 3.7, Definition 3.9 and Definition 3.11 in the literatures are stated in the remarks that follows the definitions.

Definition 3.7 (R-WAPG intermediate form)
Assume $\mu < L$ and let $(\alpha_k)_{k\geq 0}, (\rho_k)_{k\geq 0}$ given by Definition 3.4. Initialize any x_1, v_1 in $\mathbb{R}^n$. For $k \geq 1$, the algorithm generates sequence of vector iterates $(y_k, v_{k+1}, x_{k+1})_{k\geq 1}$ by the procedures:

For $k = 1, 2, \ldots$

$$
\begin{aligned}
y_k &= \left(1 + \frac{L - L\alpha_k}{L\alpha_k - \mu}\right)^{-1} \left(v_{k+1} + \left(\frac{L - L\alpha_k}{L\alpha_k - \mu}\right) x_k\right), \\
x_{k+1} &= y_k - L^{-1}\mathcal{G}_L y_k, \\
v_{k+1} &= \left(1 + \frac{\mu}{L\alpha_k - \mu}\right)^{-1} \left(v_k + \left(\frac{\mu}{L\alpha_k - \mu}\right) y_k\right) - \frac{1}{L\alpha_k}\mathcal{G}_L y_k.
\end{aligned}
$$

Remark 3.8 This form of APG is rarely identified in the literatures. The closest algorithm that fits the form but with $\mu = 0$ is Chapter 12 of in Ryu and Yin's Book [24], right after Theorem 17. We created this form which makes the math that follows simpler. The inspiration of using this as an intermediate representation was inspired by solving Exercise 12.1 in the same Ryu and Yin's Book.

Definition 3.9 (R-WAPG similar triangle form)
Given any (x_1, v_1) in $\mathbb{R}^n$. Assume $\mu < L$. Let the sequence $(\alpha_k)_{k\geq 0}, (\rho_k)_{k\geq 0}$ be given by Definition 3.4. For $k \geq 1$, the algorithm generates sequences of vector iterates $(y_k, v_{k+1}, x_{k+1})_{k\geq 1}$ by the procedures:

For $k = 1, 2, \ldots$

$$
\begin{aligned}
y_k &= \left(1 + \frac{L - L\alpha_k}{L\alpha_k - \mu}\right)^{-1} \left(v_k + \left(\frac{L - L\alpha_k}{L\alpha_k - \mu}\right) x_k\right), \\
x_{k+1} &= y_k - L^{-1}\mathcal{G}_L y_k, \\
v_{k+1} &= x_{k+1} + (\alpha_k^{-1} - 1)(x_{k+1} - x_k).
\end{aligned}
$$

Remark 3.10 The word similar triangle form can be traced back to several literatures. The term "Method of Similar Triangle" was used for Algorithm (6.1.19) in Nesterov's book [20], but without the necessary graphical illustrations to clarify it. Equation (2), (3), (4) in [5] is a similar triangle formulation of FISTA with $\mu = 0$. To see graphical visualization on why such term is used to describe the APG algorithm in the literatures, see (3.1, 4.1) in Lee et al. [13] and Ahn and Sra [1].

Definition 3.11 (R-WAPG momentum form) *Given any $y_1 = x_1 \in \mathbb{R}^n$, and sequences*

$(\rho_k)_{k\geq 0}, (\alpha_k)_{k\geq 0}$ *Definition 3.4. The algorithm generates iterates* x_{k+1}, y_{k+1} *For* $k = 1, 2, \cdots$ *by the procedures:*

For $k = 1, 2, \ldots$

$$x_{k+1} = y_k - L^{-1}\mathcal{G}_L y_k,$$
$$y_{k+1} = x_{k+1} + \frac{\rho_k \alpha_k (1-\alpha_k)}{\rho_k \alpha_k^2 + \alpha_{k+1}}(x_{k+1} - x_k).$$

In the special case where $\mu = 0$*, the momentum term can be represented without relaxation parameter* ρ_k*:*

$$(\forall k \geq 1) \quad \frac{\rho_k \alpha_k (1-\alpha_k)}{\rho_k \alpha_k^2 + \alpha_{k+1}} = \alpha_{k+1}(\alpha_k^{-1} - 1).$$

Remark 3.12 This format fits with (2.2.19) in Nesterov's book [20], however, the sequence $(\alpha_k)_{k\geq 0}$ would be given by a different rule. See Theorem 4.4 and Lemma 4.1 to see a specific choice of $(\alpha_k)_{k\geq 0}, (\rho_k)_{k\geq 0}$ such this equivalent form of R-WAPG is in fact two possible variants of the FISTA algorithm.

4 R-WAPG unifies existing accelerations scheme

In addition to various equivalent forms of the R-WAPG algorithm, the R-WAPG sequences are much more flexible. They generalize many existing sequences used in accelerated proximal gradient schemes.

This section demonstrates that several variants of FISTA in the literatures reduces to the R-WAPG method by setting up the R-WAPG sequences with additional assumptions. This section will also demonstrate that the convergence claim (Proposition 3.6) holds, and it derives convergence rates consistent with results in the literatures. The table below shows the R-WAPG convergence results specialized into various settings.

Algorithm	μ	α_k	ρ_k	Convergence of $F(x_k) - F^*$
R-WAPG in Definition 3.5	$\mu \geq 0$	$\alpha_k \in (\mu/L, 1)$	$\rho_k > 0$	$\mathcal{O}\left(\prod_{i=0}^{k-1} \max(1, \rho_i)(1-\alpha_{i+1})\right)$ (Proposition 3.6)
Chambolle, Dossal 2015 [5]	$\mu = 0$	$0 < \alpha_k^{-2} \leq \alpha_{k+1}^{-1} - \alpha_{k+1}^{-2}$	$\rho_k \geq 1$	$\mathcal{O}(\alpha_k^2)$ (Theorem 4.4)
V-FISTA Beck (10.7.7) [2]	$\mu > 0$	$\alpha_k = \sqrt{\mu/L}$	$\rho_k = 1$	$\mathcal{O}\left((1-\sqrt{\mu/L})^k\right)$, (Theorem 4.6, remark)
R-WAPG in Definition 3.5	$\mu > 0$	$\alpha_k = \alpha \in (\mu/L, 1)$	$\rho_k = \rho > 0$	$\mathcal{O}\left(\max(1-\alpha, 1-\mu/(\alpha L))^k\right)$ (Theorem 4.6)

The lemma follows characterizes momentum sequences in the literatures using Definition 3.4.

Lemma 4.1 (R-WAPG sequences as inverted FISTA sequence) *Let R-WAPG sequence $(\rho_k)_{k\geq 0}, (\alpha_k)_{k\geq 0}$ given by Definition 3.4. If $\mu = 0, \rho_k \geq 1 \ \forall k \geq 0$, and $\alpha_0 = 1$, then:*

(i) $\alpha_k^{-2} \geq \alpha_{k+1}^{-2} - \alpha_{k+1}^{-1} \ \forall k \geq 0$

(ii) *Let $t_k := \alpha_k^{-1}$, then $0 < t_{k+1} \leq (1/2)\left(1 + \sqrt{1 + 4t_k^2}\right) \ \forall k \geq 0$, hence the name: "Inverted FISTA sequence".*

(iii) $\prod_{i=1}^{k} \max(1, \rho_{k-1})(1 - \alpha_k) = \alpha_k^2 \quad (\forall k \geq 1)$.

Remark 4.2 The sequence t_k is exactly the same as in Theorem 3.1 of Chambolle, Dossal [5].

Lemma 4.3 (Constant R-WAPG sequences) *Suppose $(\alpha_k)_{k\geq 0}, (\rho_k)_{k\geq 0}$ are R-WAPG sequences given by Definition 3.4 and assume $L > \mu > 0$. Define $q := \mu/L$. Then $\forall r \in \left(\sqrt{q}, \sqrt{q^{-1}}\right)$, the constant sequence $\alpha_k := r\sqrt{q}$ has the following:*

(i) *For any $r \in \left(\sqrt{q}, \sqrt{q^{-1}}\right)$, the constant sequence $\alpha_k := \alpha \in (q, 1)$ and $\rho_k := \rho = \left(1 - r^{-1}\sqrt{q}\right)\left(1 - r\sqrt{q}\right)^{-1} > 0$, hence it's a pair of valid R-WAPG sequence.*

(ii) *The momentum terms θ_{t+1} in Definition 3.11, which we denoted by θ is the constant: $\theta = (1 - r^{-1}\sqrt{q})(1 - r\sqrt{q})(1 - q)^{-1}$*

(iii) *When $r = 1$, $\theta = (1 - \sqrt{q})(1 + \sqrt{q})^{-1}$.*

(iv) *For all $r \in \left(1, \sqrt{q^{-1}}\right)$, $\rho > 1$; for all $r \in \left(\sqrt{q}, 1\right]$ $\rho \leq 1$.*

(v) *For all $r \in \left(\sqrt{q}, \sqrt{q^{-1}}\right)$, $\max(\rho, 1)(1 - \alpha) = \max\left(1 - r\sqrt{q}, 1 - r^{-1}q\right)$.*

Theorem 4.4 (FISTA first variant Chambolle, Dossal 2015)
Fix arbitrary $a \geq 2$. Define $\forall k \geq 1$ the sequence $(\alpha_k)_{k\geq 0}, (\rho_k)_{k\geq 0}$ by

$$\alpha_k = a/(k + a),$$
$$\rho_k = \frac{(k + a)^2}{(k + 1)(k + a + 1)}.$$

Consider the algorithm given by:

Initialize any $y_1 = x_1$.
For $k = 1, 2, \ldots$, *update:*

$$\begin{aligned} x_{k+1} &:= y_k + L^{-1}\mathcal{G}_L(y_k), \\ \theta_{t+1} &:= \alpha_{k+1}(\alpha_k^{-1} - 1), \\ y_{k+1} &:= x_{k+1} + \theta_{k+1}(x_{k+1} - x_k). \end{aligned}$$

If $\mu = 0$, *then* $(\alpha_k)_{k\geq 0}, (\rho_k)_{k\geq 0}$ *is a valid pair of R-WAPG sequence from Definition 3.4 and the above algorithm is a valid form of R-WAPG.*

Assume minimizer x^* *exists for function* F. *Then algorithm produces* $(x_k)_{k\geq 0}$ *such that* $F(x) - F(x^*)$ *convergences at a rate of* $\mathcal{O}(\alpha_k^2)$.

Remark 4.5 This algorithm described here is exactly the same algorithm being analyzed in the paper by Chambolle, Dossal [5].

Theorem 4.6 (Fixed momentum APG) *Assume* $L > \mu > 0$, *let a pair of constant R-WAPG sequence:* $(\alpha_k)_{k\geq 0}, (\rho_k)_{k\geq 0}$ *be given by Lemma 4.3. Define* $q := \mu/L$ *and for any fixed* $r \in \left(\sqrt{q}, \sqrt{q^{-1}}\right)$, *let* $\alpha_k := \alpha = r\sqrt{q}$ *be the constant R-WAPG sequence. Consider the algorithm with a constant momentum specified by the following:*

Define $\theta = \left(1 - r^{-1}\sqrt{q}\right)(1 - r\sqrt{q})(1 - q)^{-1}$.
Initialize $y_1 = x_1$; *for* $k = 1, 2, \ldots, N$, *update:*

$$\begin{aligned} x_{k+1} &= y_k + L^{-1}\mathcal{G}_L y_k, \\ y_{k+1} &= x_{k+1} + \theta(x_{k+1} - x_k). \end{aligned}$$

Then the algorithm generates $(x_k)_{k\geq 1}$ *such that* $F(x) - F(x^*)$ *convergences at a rate of* $\mathcal{O}\left(\max(1 - r\sqrt{q}, 1 - r^{-1}\sqrt{q})^k\right)$.

Remark 4.7 When $r = 1$, the algorithm described above is exactly the same as the V-FISTA algorithm specified in (10.7.7) of Beck's book [2].

5 The method of Free R-WAPG

This section introduces an algorithm of our creation inspired by the remark of Proposition 3.3. Algorithm 1 estimates the μ constant as the algorithm executes and pools the information using the Bregman Divergence of the smooth part function f.

Algorithm 1 Free R-WAPG

1: **Input:** $f, g, x_0, L > \mu \geq 0, \in \mathbb{R}^n, N \in \mathbb{N}$
2: **Initialize:** $y_0 := x_0; L := 1; \mu := 1/2; \alpha_0 = 1;$
3: **Compute:** $f(y_k);$
4: **for** $k = 0, 1, 2, \cdots, N$ **do**
5: **Compute:** $\nabla f(y_k); x^+ := [I + L^{-1}\partial g](y_k - L^{-1}\nabla f(y_k));$
6: **while** $L/2\|x^+ - y\|^2 < D_f(x^+, y)$ **do**
7: $L := 2L;$
8: $x^+ = [I + L^{-1}\partial g](y_k - L^{-1}\nabla f(y_k));$
9: **end while**
10: $x_{k+1} := x^+;$
11: $\alpha_{k+1} := (1/2)\left(\mu/L - \alpha_k^2 + \sqrt{(\mu/L - \alpha_k^2)^2 + 4\alpha_k^2}\right);$
12: $\theta_{k+1} := \alpha_k(1 - \alpha_k)/(\alpha_k^2 + \alpha_{k+1});$
13: $y_{k+1} := x_{k+1} + \theta_{k+1}(x_{k+1} - x_k);$
14: **Compute:** $f(y_{k+1})$
15: $\mu := (1/2)(2D_f(y_{k+1}, y_k)/\|y_{k+1} - y_k\|^2) + (1/2)\mu;$
16: **end for**

Line 5-8 estimates upper bound for the Lipschitz constant and find x^+, the next iterates produced by proximal gradient descent on previous y_k; Line 9 updates x_{k+1} to be x^+, a successful iterate identified by the Lipschitz line search routine; Line 10 updates the R-WAPG sequence α_k for the iterates y_{k+1}; Line 13 updates μ using the Bregman Divergence of f from iterates y_{k+1}, y_k.

Assume L given is an upper bound of the Lipschitz smoothness constant of f, then the algorithm calls $f(\cdot)$ two times, and $\nabla f(\cdot)$ once per iteration. The algorithm computes $\nabla f(y_k)$ once for x^+, $f(y_{k+1})$ once for Bregman Divergence because $f(y_k)$ is evaluated from the previous iteration, and $f(x^+)$ once for Lipschitz constant line search condition. We note that $f(y_0)$ is computed before the start of the for loop. And finally, it evaluates proximal of g at $y_k - L^{-1}\nabla f(y_k)$ once.

5.1 Numerical experiments

This section gives figures and visual for numerical experiments conducted on the R-WAPG algorithm, and other algorithms in the literatures such as the V-FISTA, and M-FISTA algorithm. We implemented and compare V-FISTA, M-FISTA from Beck, and Algorithm 1 given this section. The results of the experiments are visualized and the setup of the numerical experiments are described in the sections that follows.

The equivalences highlighted in Proposition **??** allows us to compare the sequence of iterates $(x_k)_{k\geq 1}, (y_k)_{k\geq 0}$ for R-WAPG, VISTA and M-FISTA.

Given the same randomized initial condition for all the algorithm, we measure the aggregate statistics of the base two logarithms of the normalized optimality gap (NOG), at each iteration k. Given the iterates x_k, and the minimum F^*, the normalized optimality gap we defined is:

$$\delta_k := \log_2 \left(\mathbf{NOG}_k := \frac{F(x_k) - F^*}{F(x_0) - F^*} \right).$$

Since it's not the case that F^* is always known in prior, we used the minimum of all $F(x_k)$ across all algorithms, all iterations k as the surrogate for F^*.

For the termination conditions of the algorithm, we consider the norm of the gradient mapping $\|\mathcal{G}_L(y_k)\| < \epsilon$. The L can change during each iteration if it's obtained through the specified Lipschitz line search routine.

5.1.1 Simple convex quadratic

Consider the minimization problem of $\min_x\{F(x) := f(x) + 0\}$ where the objective function is given by:

$$F(x) = (1/2)\langle x, Ax\rangle.$$

The matrix A is set to be positive semi-definite and diagonal. Then the optimization problem admits unique minimizer $x^* = \mathbf{0}$ and the minimum is zero.

We apply Algorithm 1, M-FISTA, and V-FISTA. The parameters for setting up the problem now follows.

(i) N, the dimension of the problem.

(ii) $0 < \mu < L$, the strong convexity and Lipschitz smoothness constant. They are given in prior to construct the problem.

(iii) $A \in \mathbb{R}^{N\times N}$, a diagonal matrix given by $N-1$ linearly spaced with equal increment on the interval $[\mu, L]$, and an extra number 0, i.e: $A = \mathrm{diag}(0, \mu + (L-\mu)(N-1)^{-1}, \mu + 2(L-\mu)(N-1)^{-1}, \cdots, \mu + (N-2)(L-\mu)^{-1}, L)$.

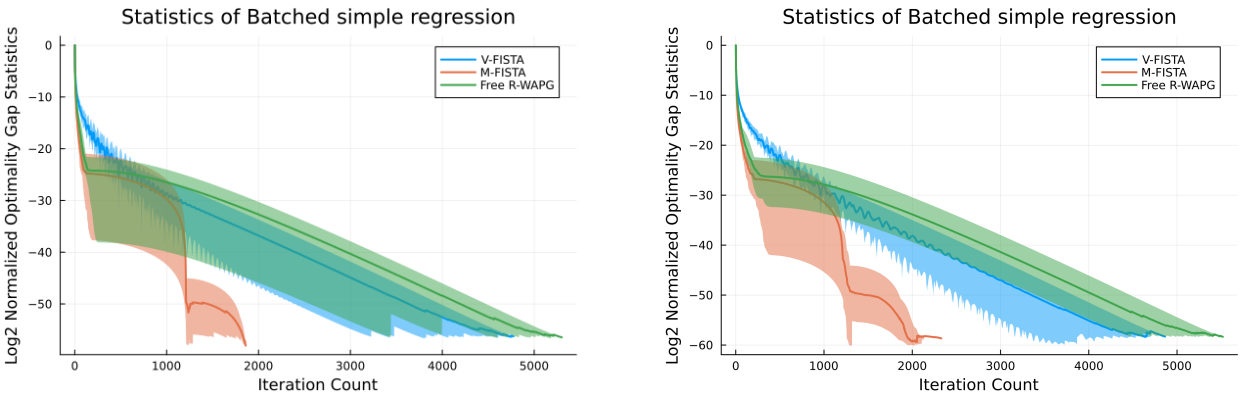
(iv) In this case $f = F = (1/2)\langle x, Ax\rangle$ and $g \equiv 0$.

(v) $\epsilon > 0$, the tolerance value for termination criteria.

(vi) $x_0 \sim \mathcal{N}(I, \mathbf{0})$ is a vector, and it's the initial condition for all the algorithm. In this case the initial guess is fixed for all R-WAPG, M-FISTA and M-FISTA, but it's randomly generated by the zero mean standard normal distribution for each element in the vector.

The parameter $L = 1, \mu = 10^{-5}$ are given in prior to produce the diagonal matrix A, and we conduct many experiments for $N = 256$ and $N = 1024$. For all R-WAPG, M-FISTA and V-FISTA, we use a different initial guess each time, a set of 30 experiments are performed. The maximum, minimum and median values of δ_k are measured for all algorithm at each iteration and plotted as a ribbon. Results are shown in Figure 1. The solid line in the ribbon is the median value of δ_k across all experiment, the ribbon gives the maximum, minimum value of δ_k for each iteration across all experiments. R-WAPG initially behaves similar to M-FISTA, but as the iteration goes on, it started to behave like V-FISTA.

The most surprising feature here is the monotone descent, however, it's being numerical verified that the method is not monotone in general, it just looks monotone on the figure.



(a) $N = 256$, simple convex quadratic.

(b) $N = 1024$, simple convex quadratic.

Figure 1: Simple convex quadratic experiments results for V-FISTA, M-FISTA, and R-WAPG.

Another quantity that maybe interesting other than δ_k would be the estimated value of μ during at each iteration k. This μ parameter should converge to the true value. One individual experiment is carried out for the R-WAPG algorithm and the value of μ at each iteration is being recorded as well. Figure 2 showcases the results. The values oscillate and converges to the true μ value. Observe that the iteration when the estimates are nearing the true value corresponds to the iteration when the algorithm plateau away from its initial fast descent.

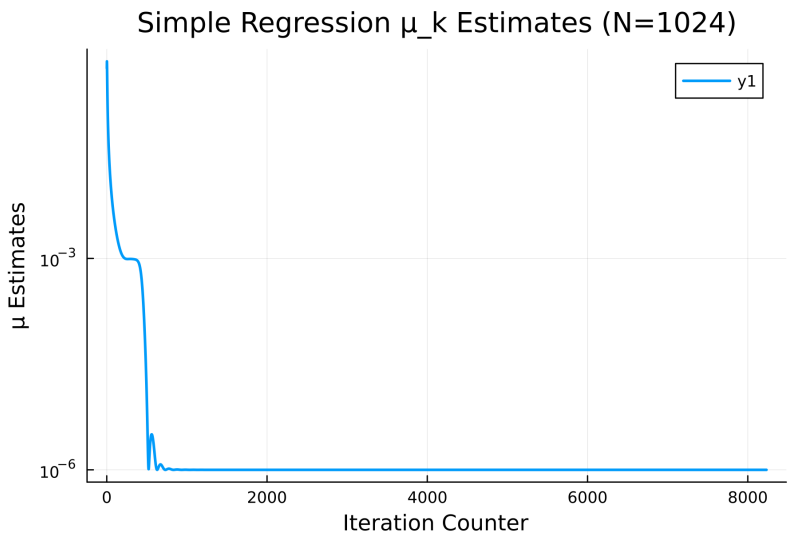


Figure 2: $N = 1024$, the μ estimates produced by Algorithm 1 (R-WAPG) is recorded.

5.1.2 LASSO

This section present results of numerical experiment for solving the (least absolute shrinkage and selection operator) LASSO problem proposed by Tibshirani [27]. The problem of LASSO has smooth, nonsmooth additive and the problem is given by:

$$\min_x \left\{ \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1 \right\}.$$

The smooth part is $f = \frac{1}{2}\|Ax-b\|^2$ and the nonsmooth is $g = \lambda\|x\|_1$. The objective function is coercive and the exact minimum, or minimizers are unknown. We perform numerical experiments using V-FISTA, M-FISTA and R-WAPG on this problem. The parameters for setting up the problem now follow.

(i) M, N are constants.

(ii) $A \in \mathbb{R}^{M\times N}$ is a matrix of i.i.d random variable, taken from a standard normal distribution.

(iii) L, μ, the Lipschitz constant and the strong convexity constant for the smooth part of the objective are not known prior, and it's estimated through A by $\mu = 1/\|(A^TA)^{-1}\|$ and $L = \|A^TA\|$.

(iv) $x^+ = [1\ -1\ 1\ \cdots]^T \in \mathbb{R}^N$, it's a vector with alternating $1, -1$ in it.

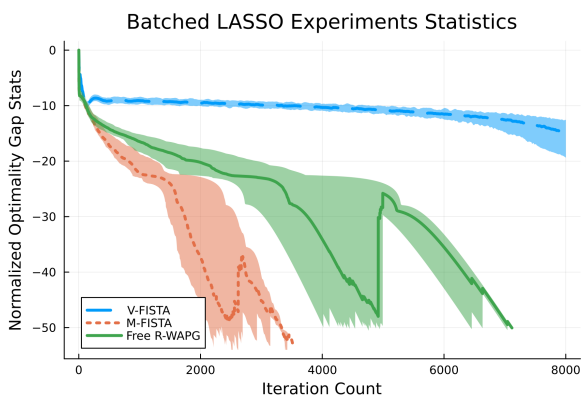
(v) Given x^+, it has $b = Ax^+ \in \mathbb{R}^M$.

(vi) Given A, estimations for L, μ are given by $L = \|A^T A\|$, $\mu = \|(A^T A)^{-1}\|^{-1}$.

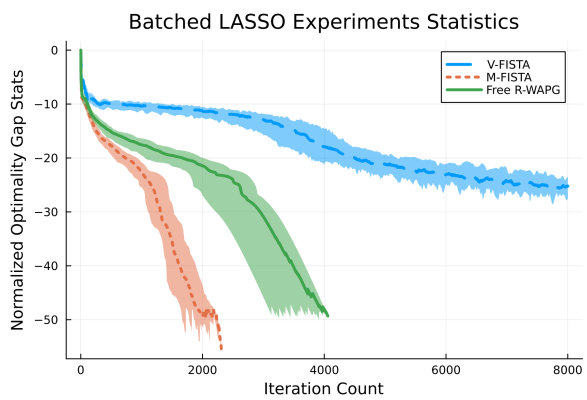
(vii) $x_0 \in \mathbb{R}^N$ is the initial guess. Its elements are random i.i.d variable realized from the standard normal distribution.

(viii) $\epsilon > 0$ is the tolerance the controls the termination criteria for test algorithms.

Experiments were conducted using V-FISTA, M-FISTA and R-WAPG with $(M, N) = (64, 256)$ and $(M, N) = (64, 128)$. Matrix A is fixed and the for all test algorithms and all repetitions. The same experiment are repeated 30 times, but each time, we fix a different random initial condition x_k for all test algorithms. The aggregate statistics of δ_k are collected for all repetitions, and then grouped by the respective algorithm. The results are showcased in Figure 3. The bump on the curve is due to a subset of test instances of the 30 repetition where the algorithms take larger number of iterations to terminate.



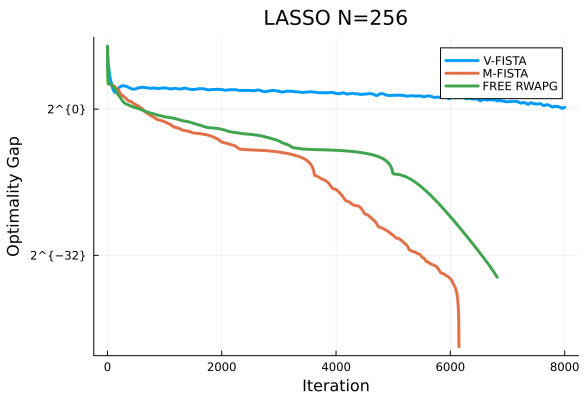
(a) LASSO experiment with $M = 64, N = 256$. Plots of minimum, maximum, and median δ_k with estimated F^*.



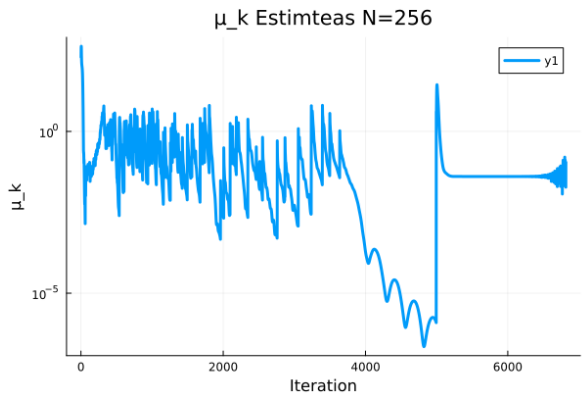
(b) LASSO experiment with $M = 64, N = 128$. Plots of minimum, maximum, and median δ_k with estimated F^*.

Figure 3: LASSO experiments.

Another quantity of interest is the estimates of μ on each iteration of the algorithm. A single experiment were conducted and the estimates and δ_k are showcased in Figure 4



(a) Single lasso experiment plot of δ_k with.



(b) The μ estimated by test algorithms for one LASSO experiment.

Figure 4: A single LASSO experiment results, with $M = 64, 256$.

For this specific experiment showed in the figure, the estimated value of μ, L which we feed into V-FISTA are $\mu = 7.432363627613958 \times 10^{-18}$ and $L = 2321.737206983643$. One of the most important feature is that the estimate μ doesn't converge to the true value, but it didn't affect the convergence of δ_k.

5.2 Future works for R-WAPG

The R-WAPG framework neglected a detail in the literatures. Our future works would extend the frameworks and include this detail.

5.2.1 Nesterov's idea of strong convexity transfer

We consider the case that $F = f + g$ where, f is L Lipschitz smooth, $\mu_f \geq 0$ strongly convex and g is closed convex. In Nesterov's 2013 paper [18], he considers accelerated minimization problem $\phi = f + \Psi$ with f L_f smooth and Ψ being $\mu_\Psi \geq 0$ strongly convex.

This detail on itself is not necessarily interesting, since without lost of generality, we can always do the splitting $f + \mu_\Psi/2\|\cdot\|^2$ and $\Psi - \mu_\Psi/2\|\cdot\|^2$ instead so that the smooth part is $\mu_\Psi \geq 0$ strongly convex. It's an interesting detail we should consider because:

(i) A strongly convex nonsmooth parts still gives linear convergence by Theorem 6 in Nesterov's 2013 [18].

(ii) The strong convexity constant can be easily estimated via a routine specified as (5.14) in Nesterov 2013 [18] and the complexity of the estimation is bounded precisely.

(i) indicates that our theories of R-WAPG is incomplete, since it doesn't predict linear convergence when g is strongly convex. (ii) indicates that splitting the strongly convex objective so that it's with the proximal operator exposes computational advantage.

Furthermore, Nesterov was not the only person who thought about it. In Chambolle, Pock [6] Algorithm 5, several variants of the FISTA algorithm were captured together into one formulation, and it assumes that $F = f + g$ where f, g has strong convexity index $\mu_f \geq 0, \mu_g \geq 0$, and F is $\mu := \mu_f + \mu_g \geq 0$ strongly convex.

6 Catalyst accelerations and future works

Assumption 6.1 Given any $\beta > 0$ and $y \in \mathbb{R}^n$, and $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ is $\mu \geq 0$ strongly convex and closed. Assume that minimizer exists for F and the minimum is F^*. Define the model function for all $y \in \mathbb{R}^n$ to be

$$\mathcal{M}_F^{\beta^{-1}}(x; y) := F(x) + \frac{\beta}{2}\|x - y\|^2.$$

We define the Moreau Envelope at $y \in \mathbb{R}^n$ to be $\mathcal{M}^*_{F,\beta^{-1}}(y) = \min_{x \in \mathbb{R}^n} \mathcal{M}_F^{\beta^{-1}}(x; y)$. We denote $\mathcal{J}_{\beta^{-1}F}$ to be the resolvent operator for subgradient of F.

Definition 6.2 (Absolute termination criterion C1) *Take F as given by Assumption 6.1. Given any $\epsilon > 0, \kappa > 0$ and $x \in \mathbb{R}^n$, the absolute criterion C1 characterizes the set of inexact proximal iterates as the set:*

$$\mathcal{J}^{\epsilon}_{\kappa^{-1}F}(x) := \left\{ y \in \mathbb{R}^n \;\middle|\; \mathcal{M}_F^{1/\kappa}(y; x) - \mathcal{M}^*_{F,1/\kappa}(x) \leq \epsilon \right\}.$$

Remark 6.3 Setting $\epsilon = 0$, we have the exact definition of the exact resolvent given as $\mathcal{J}_{\beta^{-1}} y = \mathcal{J}^0_{\beta^{-1}} y$.

Organizations now follow. Section 6.1 gives the main ideas behind the Catalyst Acceleration Framework. It follows Lin et al.'s first paper [14] on Catalyst Acceleration, the key innovations and what it can be used for together with a sketch over the complexity analysis of the entire frameworks. Section 6.2 introduces Lin et al.'s second paper [15] on Catalyst which considers an improved relative error condition stated in 6.15 that bounds the inexactness of the gradient on the Moreau Envelope. This condition not only improves the overall complexity, but it also hints the development of nonconvex gradient based 4WD Catalyst by Paquette [22].

6.1 Introduction to Catalyst

Inspired by accelerated proximal point method from Güler [12], and inexact proximal point method of Rockafellar 1976 [23], Lin [14] proposed a generic method taking inspirations from the convergence claims of Accelerated proximal point method to accelerated the convergence rate of first order variance reduced incremental method. The class of variance reduced method is vast, but to use the most relevant feature of this class of first order method is that they are stochastic method that is not slower than full gradient descent in complexity. See Gower's guide [10] for more information on variance reduced methods in machine learning.

In brief, a variance reduce method (VRM) is a type of incremental methods for solving a large sum problem: $F(x) = \sum_{i=1}^{N} f_i(x)$ in machine learning. See Bertsekas's surveys [3, 4] on incremental method for more context. VRM can be deterministic, or stochastic. When it's stochastic, the theories focuses on the expected optimality gap: $\mathbb{E}[F(x_k) - F^*]$, for the inner and outer loop. Let's assume for simplicity of discussion that it's deterministic and focus on: $F(x_k) - F^*$.

The idea of VRM is to stabilize the estimate of gradient using information of the gradient (which can be estimated, or exact) from all or a subset at previous iterates. In each iteration, the gradient of just a few samples are used to attain a new estimate of the gradient with minimum additional calculations. It gives better complexity than full gradient descent. Compare to traditional stochastic gradient, a smaller variance of the estimated gradients near the minimizer gives faster convergence rate. Major examples of VRMs include SVRG by Xiao, Zhang [28], Finito by Defazio et al. [8], SAG by Schmidt et al. [25], and SAGA by [7].

The coming parts introduce the Catalyst algorithms and its key innovations. The last section reviews recent literatures on the topics and gives potential future works.

Definition 6.4 (Lin's Universal Catalyst Acceleration)
Let $F : \mathbb{R} \to \overline{\mathbb{R}}$ be $\mu \geq 0$ strongly convex and closed. Let the initial estimate be $x_0 \in \mathbb{R}^n$, fix parameters $\kappa > 0$ and $\alpha_0 \in (0, 1]$. Let $(\epsilon_k)_{k\geq 0}$ be an error sequence chosen for the evaluation for inexact proximal point method.

Initialize $x_0 = y_0$. Then the algorithm generates $(x_k, y_k)_{k\geq 0}$ for all $k \geq 1$ such that:

$$
\begin{aligned}
&\text{find } x_k \in \mathcal{J}^{\epsilon_k}_{\kappa^{-1}F} y_{k-1}, \\
&\text{find } \alpha_k \in (0,1) \text{ such that } \alpha_k^2 = (1-\alpha_k)\alpha_{k-1}^2 + (\mu/(\mu+\kappa))\alpha_k, \\
&y_k = x_k + \frac{\alpha_{k-1}(1-\alpha_{k-1})}{\alpha_{k-1}^2 + \alpha_k}(x_k - x_{k-1}).
\end{aligned}
$$

Remark 6.5 The above algorithm is Algorithm 1 from the first paper on Catalyst Acceleration by Lin et al. [14]. The explicit formula for α_k is the larger root of solving the quadratic equation given by:

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

where $q = \mu/(k+\mu)$. Lin suggests different choices for the parameter $\kappa > 0$ depending on the algorithm chosen to evaluate the subroutine for $\mathcal{J}_{\kappa^{-1}F}^{\epsilon_k} y_{k-1}$. The choice of ϵ_k depends on the estimated optimality gap $F(x_0) - F^*$ where F^* is the minimum of F and whether $\mu > 0$ or $\mu = 0$.

Notations now follows. With a fixed regularization parameter $\kappa > 0$, the outer loop (Definition 6.4) produces $(y_k, x_k)_{k \geq k}$ denote it by $\mathbb{A}$. Typically, an iterative scheme (i.e: VRM) with a known complexity is assigned to find inexact proximal iterates $x_k \in \mathcal{J}_{\kappa^{-1}F}^{\epsilon_k} y_{k-1}$. We refer this algorithm as the inner loop and denote it by $\mathbb{M}$. Without loss of generality assume it generates some convergence sequence $(z_{k,t})_{t \geq 0}$ where k is the corresponding iteration counter of the outer loop.

The choice of error sequence $(\epsilon_k)_{k \geq 0}$ determines iteration $\mathbb{A}, \mathbb{M}$. The total iteration complexity of the algorithm counts the total number of inner loop iteration. The term "iteration complexity" refers to the number of iterations required to achieve a desired accuracy, a concept related to the convergence rate of the algorithm. The term "complexity" refers to the total number of oracle calls, in our context it depends on the specific implementation of $\mathbb{M}$. Due to the fact that function F is convex, we focus on the convergence rate of the optimality gap $F(x_k) - F^*$ for $\mathbb{A}$ and convergence of the model function $\mathcal{M}_F^{\kappa^{-1}}(\cdot, y_{k-1})$ for $\mathbb{M}$ given y_{k-1}, ϵ_k and initial guess $z_{k,0}$.

6.1.1 Outer loop iteration complexity

The error sequence $(\epsilon_k)_{k \geq 0}$ governs the convergence rate of the outer loop for $F(x_k) - F^*$ to converge. Depending on either $\mu > 0$, or $\mu = 0$, the choice of $(\epsilon_k)_{k \geq 0}$ differs. The theorems that follow state the error sequence required for the outer loop to retain optimal convergence rate, they are Theorem 3.3, 3.1 in Lin et al. [14].

Theorem 6.6 (Outer loop convergence strongly convex) *For $\mathbb{A}$ with regularization parameter $\kappa > 0$. Assume that F is $\mu > 0$ strongly convex. Choose $\alpha_0 = \sqrt{q}$ with $q = \mu/(\kappa + \mu)$ and the error sequence*

$$\epsilon_k = \frac{2}{9}(F(x_0) - F^*)(1-\rho)^k \quad \textit{with} \quad \rho < \sqrt{q}.$$

Then the $\mathbb{A}$ *generates* $(x_k)_{k\geq 0}$ *such that*

$$F(x_k) - F^* \leq C(1-\rho)^{k+1}(F(x_0) - F^*) \quad \textit{with} \quad C = \frac{8}{(\sqrt{q}-\rho)^2}.$$

Remark 6.7 Suggested by Lin et al., ρ is at the discretion of the practitioner, take for example $\rho = 0.9\sqrt{q}$ would work.

Theorem 6.8 (Outer loop convergence convex but not strongly convex) *For* $\mathbb{A}$ *with regularization parameter* $\kappa > 0$*. Assume that* F *is convex but with strong convexity constant* $\mu = 0$*. Choose* $\alpha_0 = (\sqrt{5}-1)/2$ *and the error sequence*

$$\epsilon_k = \frac{2(F(x_0) - F^*)}{9(k+2)^{4+\eta}} \quad \textit{with} \quad \eta > 0.$$

Take x^* *to be a minimizer of* F*. Then algorithm* $\mathbb{A}$ *generates* $(x_k)_{k\geq 0}$ *such that it has a convergence rate of*

$$F(x_k) - F^* \leq \frac{8}{(k+2)^2}\left(\left(1+\frac{2}{\eta}\right)^2 (F(x_k) - F^*) + \frac{\kappa}{2}\|x_0 - x^*\|^2\right).$$

Remark 6.9 Suggested by Lin et al., $\eta > 0$ is at the discretion of the practitioners, for an example, $\eta = 0.1$ would work.

The same convergence claims hold if smaller value of ϵ_k is taken, Theorem 6.6, 6.8 gives the largest possible error sequence $(\epsilon_k)_{k\geq 0}$ such that the convergence claim holds. Of course, taking smaller values of $(\epsilon_k)_{k\geq 0}$ as suggested by the above theorems for $\mathbb{M}$ will hinder its convergence.

Observe that the error sequence $(\epsilon_k)_{k\geq 0}$ requires prior knowledge of F^*. It's poses no theoretical concerns, but it's of upmost practical concern since F^* may not be accessible in practice prior to executing the algorithm. In the work by Lin et al., the example algorithm given is D.3 Accelerating MISO Prox. It automatically builds a lower bound estimates on F^* in the outer loop as the algorithm executes.

The convergence of the outer loop made use of an inexact version of the proximal gradient inequality (similar to Theorem 2.16) stated as Lemma A.7 in [14]. This lemma is instrumental for deriving an inexact variant of the estimating sequence $\phi_k^* \geq F(x_k) + \xi_k$. The convergence proof (outer and inner loop together) from Lin was inspired by Schmidt's Inexact Proximal Gradient method [26]. The technique of estimating sequence introduced back in Definition 2.18 did the heavy lifting, but it results in depressingly long proof making it unsuitable for exposition here. Significant pieces of theoretical innovations are covered in details in our most recent Fall Winter 2024 MATH 590 report. The parts that come will complement content in the report, filling up missing content and attempts to build a bigger picture of Catalyst Acceleration.

6.1.2 Inner loop complexity

The iteration complexity of M relates to the outer loop when warm start is used. The Catalyst Paper [14] suggested the use of $z_{k,0} = x_{k-1}$ as the warm start condition. With Assumption 6.10, and the suggested warm start condition, Lin et al. derived the upper bound of the iteration complexity of $\mathbb{M}$, which are stated in Proposition 3.2, 3.3 in their text, and 6.14, 6.12 restates things below in our notations. These two theorems relate the convergence of $\mathbb{M}, \mathbb{A}$ and gives a convergence rate of $F(x_k) - F^*$ expressed using the total number of iteration underwent by $\mathbb{M}$.

Assumption 6.10 (Linear convergence of inner loop) Fix any $k \in \mathbb{N}$, any $y \in \mathbb{R}^n$. Suppose $\mathbb{M}$ generates iterates $(z_{k,t})_{t\geq 0}$ for the inner loop iteration such that there exists $A > 0$, and it has:

$$\mathcal{M}_F^{\kappa^{-1}}(z_{k,t}, y) - \mathcal{M}^*_{F,\kappa^{-1}}(y) \leq A(1 - \tau_{\mathbb{M}})^t \left(\mathcal{M}_F^{\kappa^{-1}}(z_{k,0}) - \mathcal{M}^*_{F,\kappa^{-1}}(y)\right).$$

Remark 6.11 Assumption is mild given the fact that model function $\mathcal{M}(\cdot, y_{k-1})$ is $\mu + \kappa$ strongly convex given Assumption 6.1. For example, with smoothness assumption on F, many VRMs, or gradient descent method can achieve linear convergence under strong convexity. Since κ is fixed, this assumption can be applied for all k for $\mathbb{A}$.

Proposition 6.12 (Inner loop complexity strongly convex) *Under the same settings of Theorem 6.6, suppose that*

(i) $\mathbb{M}$ *has linear convergence rate as specified in Assumption 6.10,*

(ii) $\mathbb{M}$ *is initialized with* $z_{k,0} = x_{k-1}$.

Then, the precision ϵ_k *is achieved within at most a number of iteration* $T_{\mathbb{M}} \leq \widetilde{\mathcal{O}}(1/\tau_{\mathbb{M}})$. *Here* $\widetilde{\mathcal{O}}$ *hides logarithmic complexity in* μ, κ *and other constants.*

Remark 6.13 $T_{\mathbb{M}}$, an upper bound of the iteration of the inner loop, is a constant in this case.

Proposition 6.14 (Inner loop, context but not strongly convex) *Under the settings of Theorem 6.8, suppose that:*

(i) $\mathbb{M}$ *has linear convergence rate as specified in Assumption 6.10,*

(ii) *the initial guess for* $\mathbb{M}$ *is* $z_{0,k} = x_{k-1}$,

(iii) F *has bounded level set.*

Then there exists $T_{\mathbb{M}} \leq \widetilde{\mathcal{O}}(1/\tau_{\mathbb{M}})$ *such that for any* $k \geq 1$. *Then it requires at most* $T_{\mathbb{M}} \log(k+2)$ *iterations for* $\mathbb{M}$ *to achieve accuracy* ϵ_k.

For a proof of Proposition 6.14, 6.12, see Appendix item B1, B2 in Lin et al. [14]. We are now ready to derive the convergence rate measured by the number of total iteration experience by $\mathbb{M}$. If $\mu > 0$, so Proposition 6.12 gives total number of inner iteration is bounded by $m \leq T_{\mathbb{M}} k$. Substituting $k \geq m/T_{\mathbb{M}}$ into Theorem 6.6, it gives description of convergence rate of the algorithm measured by the total number of iteration experience by $\mathbb{M}$:

$$\begin{aligned} F(x_k) - F^* &\leq \mathcal{O}\left((1-\rho)^k\right) \leq \mathcal{O}\left((1-\rho)^{m/T_{\mathbb{M}}}\right) \leq \mathcal{O}\left((1-\rho/T_{\mathbb{M}})^m\right) \\ &\leq \widetilde{\mathcal{O}}\left(\tau_{\mathbb{M}}\sqrt{\mu}/(\mu+\kappa)\right). \end{aligned}$$

The second inequality on the first line made use of the fact that $1 + x \leq (1 + x/n)^n$ for all $n \geq 1$ and $|x| \leq n$. The optimal value of κ is suggested by choosing the best $\kappa > 0$ that minimizes the above upper bound.

If $\mu = 0$, using Proposition 6.14 the total number of inner loop iteration executed by $\mathbb{M}$ at the k th iteration of $\mathbb{A}$ is bounded via:

$$m \leq \sum_{i-1}^{k} kT_{\mathbb{M}} \log(i+2) \leq kT_{\mathbb{M}} \log(k+2) \leq T_{\mathbb{M}} k(k+2) \leq \mathcal{O}(T_{\mathbb{M}} k^2).$$

Therefore, using Theorem 6.8, the convergence rate as measure by the total number of inner iteration is given by:

$$F(x_k) - F^* \leq \mathcal{O}(k^{-2}) \leq \mathcal{O}\left(m^{-2} T_{\mathbb{M}}\right) \leq \widetilde{\mathcal{O}}\left(m^{-2}\tau_{\mathbb{M}}^{-1}\right).$$

The last inequality is $T_{\mathbb{M}} < \widetilde{\mathcal{O}}(1/\tau_{\mathbb{M}})$ from Proposition 6.14. In both cases, the convergence rate remains optimal as measure by the total number of $\mathbb{M}$.

Lin et al.'s second paper on Catalyst Acceleration [15] describes new ideas to choose the termination criteria for the inner for evaluating $\mathcal{J}^{\epsilon_k}_{\kappa^{-1}F} y_k$ and gives alternatives to the error sequence ϵ_k based on those termination criteria. To elucidate, consider the model function $\mathcal{M}_F^{\kappa^{-1}}(x; y)$ to be $\mu + \kappa$ strongly convex. Therefore, it has error bound condition:

$$(\forall x \in \mathbb{R}^n) \quad \mathcal{M}_F^{1/\kappa}(x; y) - \mathcal{M}^*_{F,1/\kappa}(y) \leq (\kappa + \mu) \operatorname{dist}\left(\mathbf{0}, \partial\mathcal{M}_F^{1/\kappa}(x; y)\right)^2.$$

By choosing x such that $\operatorname{dist}\left(\mathbf{0}, \partial\mathcal{M}_F^{1/\kappa}(x; y)\right) \leq \sqrt{\epsilon}$, it ensures $x \in \mathcal{J}^{\epsilon}_{\kappa^{-1}F}(y)$. Unfortunately, this is a difficult in practice because a full gradient evaluation on the model function $\mathcal{M}_F^{1/\kappa}(\cdot; y)$ is costly (compare to the small amount required for VRM, which is just the gradient of a few samples.), so Lin suggested alternatives of inner loop termination criteria, and/or upper bounds of inner loop iteration to make Catalyst Acceleration competitive in practice.

6.2 The second Catalyst Acceleration paper

This section discusses major contents in Lin's second Catalyst paper [15]. To expedite the execution of $\mathbb{M}$ in general given an error sequence $(\epsilon_k)_{k\geq 0}$ (or equivalently some lower bounds of it), Lin suggested the following new ideas which are not mutually exclusive:

(i) An improved warm start strategy at the end of Section 3. It improved the convergence rate under different termination criteria, and it supports smooth plus nonsmooth objective.

(ii) A relative termination criterion C2 stated in Definition 6.15, governed by the error sequence $(\delta_k)_{k\geq 0}$. The sequence δ_k doesn't require knowledge on F^*, it simplifies the convergence proof, gives better bounds on the complexity for $\mathbb{M}$ without using bounded level set condition. Recall that Theorem 6.14 requires bounded level set assumption on F.

In our notation, we define relative termination criterion C2.

Definition 6.15 (Relative termination criterion C2) *Take F as given by Assumption 6.1. Given any $\delta \in (0,1]$, $\kappa > 0$ and $x \in \mathbb{R}^n$, the relative criterion C2 is characterized by the set:*

$$\widetilde{\mathcal{J}}^{\delta}_{\kappa^{-1}F}(x) := \left\{ z \in \mathbb{R}^n \;\middle|\; \mathcal{M}^{\kappa^{-1}}_{F}(z;x) - \mathcal{M}^{*}_{F,\kappa^{-1}}(z;x) \leq \frac{\kappa\delta}{2}\|x-z\|^2 \right\}.$$

Observe that, if we set $\epsilon_k = \delta\kappa/2\|x-z\|^2$ then $\widetilde{\mathcal{J}}^{\delta}_{\kappa^{-1}F}(x) = \mathcal{J}^{\epsilon}_{\kappa^{-1}F}(x)$. The relative inexact condition can be interpreted as an adaptive inexactness condition. Stated by Lin, the following lemmas are the sufficient conditions and consequences of termination criteria C1, C2.

Lemma 6.16 (Sufficient condition for C1) *Consider smooth plus nonsmooth objective $F := f + g$ with $g : \mathbb{R}^n \to \overline{\mathbb{R}}, F : \mathbb{R}^n \to \mathbb{R}$ closed and convex, and F is $\mu \geq 0$ strongly convex and L-Lipschitz smooth. With arbitrary $x \in \mathbb{R}^n$ fixed, the model function is additive composite of the form:*

$$\mathcal{M}^{\kappa^{-1}}_{F}(z;x) := \underbrace{f(z) + \frac{\kappa}{2}\|z-x\|^2}_{=:f_\kappa(z)} + g(z).$$

For any z define proximal gradient point

$$\bar{z} = \operatorname{prox}_{\eta g}(z - \nabla f_\kappa(z)) \quad \text{with} \quad \eta = 1/(\kappa + L).$$

Then it has

$$\frac{1}{\eta}\|z-\bar{z}\| \leq \sqrt{2\kappa\epsilon} \implies \bar{z} \in \mathcal{J}^{\epsilon}_{\kappa^{-1}F}(x).$$

Remark 6.17 This is Lemma 2 in Lin et al. [15].

6.2.1 Consequences of the inner loop termination criteria

Lemma 6.16 describes a sufficient conditions to verify the membership of $\bar{z} \in \mathcal{J}^{\epsilon}_{\kappa^{-1}F}(x)$ through the proximal gradient operator on $F := f_\kappa(z) + g(z)$ which of practical importance because it translates criterion C1 into something implementable, i.e: the proximal gradient operator. For theoretical interests, the absolute and relative criteria C1, C2 places bound on the true error of the gradient of Moreau Envelope at the point x.

The following are results proved in Lin's second Catalyst paper. Given any $\epsilon > 0$ if $z \in \mathcal{J}^{\epsilon}_{\kappa^{-1}F}(x)$, define the approximated gradient mapping $\mathcal{G}^{\epsilon}_{\kappa^{-1}F}(z) := \kappa(x-z)$. Then

$$\|z - \mathcal{J}_{\kappa^{-1}F}(x)\| \leq \sqrt{\frac{2\epsilon}{\kappa}} \iff \left\|\mathcal{G}^{\epsilon}_{\kappa^{-1}F}(z) - \nabla\mathcal{M}^{*}_{F,\kappa^{-1}}(x)\right\| \leq \sqrt{2\kappa\epsilon}.$$

Similarly, if $z \in \widetilde{\mathcal{J}}^{\delta}_{\kappa^{-1}F}(x)$, define the inexact gradient mapping $\widetilde{\mathcal{G}}^{\delta}_{\kappa^{-1}F}(x) = \kappa(x-z)$, we have:

$$\begin{aligned} \|z - \mathcal{J}_{\kappa^{-1}F}(x)\| &\leq \sqrt{\delta}\|x-z\| \leq \sqrt{\delta}(\|x - \mathcal{J}_{\kappa^{-1}F}(x)\| - \|z - \mathcal{J}_{\kappa^{-1}F}(x)\|), \\ \left\|\widetilde{\mathcal{G}}^{\delta}_{\kappa^{-1}F}(x) - \nabla\mathcal{M}^{*}_{F,\kappa^{-1}}(x)\right\| &\leq \delta'\|\nabla\mathcal{M}^{*}_{F,\kappa^{-1}}(x)\| \quad \text{with } \delta' = \sqrt{\delta}/\left(1-\sqrt{\delta}\right). \end{aligned}$$

These results are instrumental to prove the convergence of $\mathbb{M}$ and giving convergence claim of $\mathbb{A}$ under certain assumption on $(\delta_k)_{k\geq 0}, (\epsilon_k)_{k\geq 0}$. For a more general development of characterizations of inexact oracles that is more comprehensive and rigorous, see Devolder et al. [9]. It's discusses a Nesterov accelerated algorithm using estimating sequence with inexact oracles.

Besides Definition 6.2, 6.15, the "Fixed Budget" termination criterion terminates $\mathbb{M}$ after a fixed number of iteration given accuracy ϵ_k. Criterion C1 used to prove the overall complexity in Lin et al. [14], but the bound is deliberately loose to allow generality, making it extremely impractical. Criterion C2 places similar upper bound on the iteration complexity for $\mathbb{M}$, but it still remains impractical.

Interestingly, the new termination criterion C2 gives improvements on the complexity of inner loop $\mathbb{M}$, outer loop $\mathbb{A}$, and the relative error sequence $(\delta_k)_{k\geq 0}$ for $\mathbb{M}$. The theorems and commentaries that follow will illustrate.

Definition 6.18 (Catalyst Acceleration with relative inexactness) *Suppose that F is a $\mu \geq 0$ strongly smooth under Assumption 6.1. Initialize any $x_0 \in \mathbb{R}^n$, $\kappa > 0$. Given the relative error sequence $(\delta_k)_{k\geq 0}$, or equivalently some fixed budget number of iterations $T_{\mathbb{M}}$, or absolute error sequence $(\epsilon_k)_{k\geq 0}$:*

Initialize $y_0 = x_0, q = \mu/(\kappa+\mu), k = 1$, and if $\mu > 0$, set $\alpha_0 = \sqrt{q}$ otherwise $\alpha_0 = 1$.
***While** desirable accuracy is not reached, **do**:*

(i) *Finds $x_k \approx \mathcal{J}_{\kappa^{-1}F} y_{k-1}$ using warm start.*

(ii) *Pick one of the following fixed termination criterion:*

  (a) *Determine the number of fixed budget depending on $T_{\mathbb{M}}$, terminate the above subroutine if fixed budget reached.*

  (b) *C1 are satisfied through ϵ_k.*

  (c) *C2 are satisfied through δ_k.*

(iii) *Find $\alpha_k \in (0,1)$ such that $\alpha_k^2 = (1-\alpha_k)\alpha_{k-1}^2 + q\alpha_k$.*

(iv) *Compute extrapolation $y_k = x_k + \beta_k(x_k - x_{k-1})$ with*

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

(v) *Increment $k := k+1$*

End while

Theorem 6.19 (Outer loop complexity under criterion C2) *For the iterates $(x_k)_{k\geq 0}$ generated by algorithm in Definition 6.18, we have*

(i) *If $\mu > 0$, choose $\alpha_0 = \sqrt{q}$, $\delta_k = \sqrt{q}/(2-\sqrt{q})$. Then the iterates $(x_k)_{k\geq 0}$ satisfies $F(x_k) - F^* \leq \mathcal{O}\left(1 - \sqrt{q}/2\right)^k$.*

(ii) *If $\mu = 0$, choose $\alpha_0 = 1$, $\delta_k = 1/(k+1)^2$ satisfies $F(x_k) - F^* \leq \mathcal{O}(k^{-2})$.*

Remark 6.20 This is Proposition 8, 9 in Lin et al.'s second Catalyst paper [15]. For a precise description of the upper bound, see Theorem 8.

The key improvement here compared to using absolute inexactness for $\mathbb{M}$ as stated by Theorem 6.6, 6.8 is that $(\delta_k)_{k\geq 0}$ doesn't require knowledge on F^*. This innocent detail helps to develop a nonconvex variant of Catalyst call 4WD Catalyst which is the focus of the paper by Paquette [22]. Corollary 16 in Lin et al.'s second Catalyst paper [15] gives complexity

for $\mathbb{M}$ Using termination criterion C2 with a warm start specialized for $C2$. It's important to note that the upper complexity bounds are much better, and it doesn't require bounded level set compared to Theorem 6.14.

6.3 Potential future research

Now, we have enough context to understand the potential future directions of research regarding the Catalyst Acceleration framework.

6.3.1 Necoara et al.'s comments on Catalyst Acceleration

Necoara et al. [16] considered optimization problem $f^* = \min_{x \in X} f(x)$ with $f : X \subseteq \mathbb{R}^n \to \mathbb{R}$ and X closed and convex. They extended the linear convergence results of accelerated gradient method such as Nesterov's acceleration with proximal gradient into a wider setting than strong convexity.

As a secondary objective, they introduced and stated the relations between several weakened characterizations of strong convexity on a closed convex set. Definition 1, 2, 3, 4, 5, introduced the ideas of: Quasi-strong convexity, Quadratic under approximation, Quadratic gradient growth, quadratic functional growth (recall that it was stated as a consequence of strong convexity in Lemma 2.9), and Global error bound conditions.

The content of their paper is very in-depth hence for our interests, we focus on their comments on the first Catalyst paper by Lin et al. [14]. They said, and here we quote:

> It is also worth to investigate the complexity bounds obtained when wrapping linearly-convergent algorithms under the proposed conditions in an outer-loop using a generic acceleration scheme such as Catalyst [14]. This would allow to extend to non-strongly convex settings a wide range of incremental optimization algorithms designed for large finite-sum problems arising in machine learning and statistics.

We note that, their idea of extension can be applied to inner loop $\mathbb{M}$, or out loop $\mathbb{A}$, either, or both.

6.3.2 Our ideas on future works of Catalyst Acceleration

One intriguing detail in the Catalyst framework eluded us. It may not be obvious to the readers, so we shall elucidate.

The concern starts by considering the necessary assumption in VRMs. Defazio et al.[7] SAGA, Finito et al. [8] requires Lipschitz continuous gradient on each individual f_i in the objective function $F = \sum_{i=1}^{n} f_i$. Schmidt et al. [25] also requires Lipschitz smoothness for the non-smooth part of a composite objective. Xiao, Zhang, SVRG [28] requires Lipschitz gradient assumption as well. This seemingly raises a crucial question: "Where does the Lipschitz continuous gradient of the smooth part make its presence in the Catalyst Framework?"

A careful reader should realize that Assumption 6.10 encapsulates the Lipschitz smoothness assumption of VRMs. Furthermore, such an assumption is optimistic for the complexity of $\mathbb{M}$. Recall that with Lipschitz smoothness and strong convexity, a lower bound of exponential rate of convergence is established by Theorem 2.1.13 in Nesterov's book [20]. Removing the smoothness assumption for $\mathbb{M}$, the lower bound on convergence rate of any first order algorithm becomes $\mathcal{O}(1/k)$ by Theorem 3.2.5 for nonsmooth function that is strongly convex.

From a complexity theoretical point of view, Catalyst acceleration didn't attempt at narrowing the lower complexity bound bounds for functions that are convex and nonsmooth and lack of any exact proximal oracle. It is absolutely possible that, for any black box nonsmooth optimization problem, the same lower bound convergence rate $\mathcal{O}\left(1/\sqrt{k}\right)$ as stated in Theorem 3.2.1 in Nesterov's book [20] is true for Catalyst. Unfortunately it is absolutely nontrivial to see that is true because $\mathbb{A}$ optimizes Moreau Envelope $\mathcal{M}^*_{\kappa^{-1}F}(x)$ through inexact evaluation and the Moreau Envelope has Lipschitz gradient. For $\mathbb{A}$ Theorem 3.2.1 doesn't apply. Whatever the true answer is, it's not obvious. So, here is what we think is worth investigating:

> What if Assumption 6.10 is false in the Catalyst Acceleration Framework? Would there exists any kind of regularization parameters $(\kappa_k)_{k\geq 0}$, error sequence ϵ_k such that the overall complexity still retains convergence rate equals or faster than $\mathcal{O}(1/\sqrt{k})$?

Nobody has yet answered the question in the literature, but a brief searched showed that in Nesterov 2014 [19], he considers the method of accelerated gradient for function with Holder continuous gradient. If anything, there are some insight to be gained about Catalyst on this question on this work of his. It remains to be seemed why this detail is neglected when developing the Catalyst Acceleration framework. If the answer is that it can be faster than the lower bound, then we had extended the Catalyst Framework, otherwise, it leads to the unintuitive result that lower bound $\mathcal{O}(1/\sqrt{k})$ still holds.

References

[1] K. Ahn and S. Sra, *Understanding Nesterov's acceleration via proximal point method*, in Symposium on Simplicity in Algorithms, SIAM, June 2022.

[2] A. Beck, *First-order Methods in Optimization*, MOS-SIAM Series in Optimization, SIAM, israel, 2017.

[3] D. P. Bertsekas, *Incremental proximal methods for large scale convex optimization*, Mathematical Programming, 129 (2011), pp. 163–195.

[4] ——, *Incremental gradient, subgradient, and proximal methods for convex optimization: A survey*, Dec. 2017.

[5] A. Chambolle and C. Dossal, *On the convergence of the iterates of the "Fast iterative shrinkage/thresholding algorithm"*, Journal of Optimization Theory and Applications, 166 (2015), pp. 968–982.

[6] A. Chambolle and T. Pock, *An introduction to continuous optimization for imaging*, Acta Numerica, 25 (2016), pp. 161–319.

[7] A. Defazio, F. Bach, and S. Lacoste-Julien, *SAGA: A fast increment al gradient method with support for non-strongly convex composite objectives*, Dec. 2014.

[8] A. Defazio, J. Domke, and T. Caetano, *Finito: A faster, permutable incremental gradient method for big data problems*, in Proceedings of the 31st International Conference on Machine Learning, PMLR, June 2014, pp. 1125–1133.

[9] O. Devolder, F. Glineur, and Y. Nesterov, *First-order methods of smooth convex optimization with inexact oracle*, Mathematical Programming, 146 (2014), pp. 37–75.

[10] R. M. Gower, M. Schmidt, F. Bach, and P. Richtárik, *Variance-reduced methods for machine learning*, Proceedings of the IEEE, 108 (2020), pp. 1968–1983.

[11] G. N. Grapiglia and Y. Nesterov, *Accelerated regularized newton methods for minimizing composite convex functions*, SIAM Journal on Optimization, 29 (2019), pp. 77–99.

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

[13] J. Lee, C. Park, and E. Ryu, *A Geometric structure of acceleration and its role in making gradients small fast*, in Advances in Neural Information Processing Systems, vol. 34, Curran Associates, Inc., 2021, pp. 11999–12012.

[14] H. Lin, J. Mairal, and Z. Harchaoui, *A universal catalyst for first-order optimization*, in Procedings of Advances in Neural Information Processing Systems, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds., vol. 28, Curran Associates, Inc., 2015.

[15] ——, *Catalyst acceleration for first-order convex optimization: from theory to practice*, Journal of Machine Learning Research, 18 (2018), pp. 1–54.

[16] 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.

[17] Y. Nesterov, *Accelerating the cubic regularization of Newton's method on convex problems*, Mathematical Programming, 112 (2008), pp. 159–181.

[18] ——, *Gradient methods for minimizing composite functions*, Mathematical Programming, 140 (2013), pp. 125–161.

[19] Y. Nesterov, *Universal gradient methods for convex optimization problems*, Mathematical Programming, 152 (2015), pp. 381–404.

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

[21] W. Noel, *Nesterov's method for convex optimization*, SIAM Review, 65, pp. 539–562.

[22] C. Paquette, H. Lin, D. Drusvyatskiy, J. Mairal, and Z. Harchaoui, *Catalyst for gradient-based nonconvex optimization*, in Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, PMLR, Mar. 2018, pp. 613–622.

[23] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM Journal on Control and Optimization, 14 (1976), pp. 877–898.

[24] E. K. Ryu and W. Yin, *Large-scale Convex Optimization: Algorithms & Analyses via Monotone Operators*, Cambridge University Press, Cambridge, 2022.

[25] M. Schmidt, N. Le Roux, and F. Bach, *Minimizing finite sums with the stochastic average gradient*, Mathematical Programming, 162 (2017), pp. 83–112.

[26] M. Schmidt, N. L. Roux, and F. Bach, *Convergence rates of inexact proximal-gradient methods for convex optimization*, Dec. 2011. arXiv:1109.2415.

[27] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, Journal of the Royal Statistical Society. Series B (Methodological), 58 (1996), pp. 267–288.

[28] L. Xiao and T. Zhang, *A proximal stochastic gradient method with progressive variance reduction*, SIAM Journal on Optimization, 24 (2014), pp. 2057–2075.

[29] C. Ying and P. Jong-Shi, *Modern Nonconvex Nondifferentiable Optimization*, vol. 1 of MOS-SIAM Series on Optimization, MOS-SIAM, 2021.