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

Hongda Li

Department of Mathematics
University of British Columbia,
Okanagan Campus.

January 3, 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 $g : Q \to \overline{\mathbb{R}}$ 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 [15], a variant for of accelerated gradient method for strongly convex function f is discussed. We had two 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 faster 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 by our own efforts of research. Section 3, **??** are our ongoing research which present the answers to the questions.

In Section 3, we proposed the method of "Relaxed Weak Accelerated Proximal Gradient (R-WAPG)" as the foundation to describe several variants of Accelerated proximal gradient method in the literatures. The convergence theories of R-WAPG allows us to model convergence of accelerated proximal gradient method where the momentum sequence doesn't strictly follow the conditions presented in the literatures. The descriptive power of R-WAPG allows convergence analysis for all the variants using one single theorem.

In Section **??** we propose a practical algorithm that exploits a specific term in the proof of R-WAPG to achieve faster convergence for solving (1) without knowing parameter L, μ in prior. Results of numerical experiments are presented.

Section 5 are results of literatures review in MATH 590. It's based on a series of papers in the topic of Catalyst Meta Acceleration method for First Order Variance Reduced Methods. We will point out potential future direction of research of Catalyst acceleration.

■ Add Lin's papers and Paquette's papers.

Section 6, 7 preview literatures in nonsmooth optimization frontier research where progress and impacts can be made.

1.1 Theme of the research

This section specifies a theme of the research in this proposal. Out first objective is to explore the Goldilocks zones between these topics: theories of variational analysis, design of continuous optimization algorithm and applications in sciences, engineering, and statistics. Our second objective is to identify the "chemistry" occurring between properties of functions and the designs of continuous optimizations algorithm and how it impacts the convergence and behaviors of the algorithms.

2 Preliminaries

Clarify: Notations, Organizations.

This section contains the basics of contents from convex optimization, and variational analysis.

Notations.

(i) $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty, -\infty\}$

2.1 Fundamentals in non-convex analysis

Let the ambient space be $\mathbb{R}^n$ equipped with inner product and 2-norm. Let O be an open subset of $\mathbb{R}^n$, the weakest assumption we are making for the objective function $F : O \subseteq \mathbb{R}^n \to \mathbb{R}$ for 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 [19]

- Add bib
- Cite.

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 [19]

- Add bib
- Cite.

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 definition:

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

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

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:

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

This condition is stronger than differentiability. When F convex, it has descent lemma:

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

When F is convex, the converse holds. 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 any fixed $y \in \operatorname{dom}(\partial F)$, we have for all $x \in \mathbb{R}^n$:*

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

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 is another minimizer, then $F(x) \neq F(\bar{x})$, which is a direct contradiction. The quadratic growth condition over a set of minimizer is much weaker than convexity.

2.2.1 Smooth, nonsmooth additive composite

In this section, we zoom in further. Suppose that $F := f + g$ where $f : \mathbb{R}^n \to \mathbb{R}$ is convex, L Lipschitz smooth and $\mu \geq 0$ strongly convex and $g : \mathbb{R} \to \overline{\mathbb{R}}$ is convex. To make the discussion simpler, fix any $\beta \geq 0$ we define the following model functions as a $\mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$:

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

Under convexity assumption in this section, both $\widetilde{\mathcal{M}}(\cdot; y), \mathcal{M}(\cdot; y)$ is at least $\beta \geq 0$ strongly convex.

Definition 2.11 (Proximal gradient operator) *Take $F := f + g$ where $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ as defined in this section. 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, it has domain on the entire $\mathbb{R}^n$, and it's a $3/2$ averaged operator.

Definition 2.13 (Gradient mapping operator) *Take $F := f + g$ as defined in this section. 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)
Take $\widetilde{\mathcal{M}}^{L^{-1}}, \mathcal{M}^{L^{-1}}$ as defined in this section, we will have for all $x \in \mathbb{R}^n$ that:

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

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

Remark 2.16 This lemma still holds for non-convex f under prox-boundedness and weak convexity and differentiability of f.

Lemma 2.17 (The proximal gradient inequality) *Take $F := f + g$ as defined in this section. Fix any $y \in \mathbb{R}^n$, then for all x, the proximal gradient inequality is true:*

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

Remark 2.18 This lemma is proved in our draft paper.

2.3 Nesterov's estimating sequence technique

Do the following:

(i) What is Nesterov's estimating sequence.

(ii) How is it used to derive the algorithm and convergence rate of algorithm.

(iii) Where is it used and why is it important here.

Examples

(i) Example estimating sequence.

The method of Nesterov's estimating sequence for accelerated gradient method, and their nonsmooth counterparts assumes a convex function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$. The estimating sequence is a technique searching for candidate algorithm with extrapolated momentum, and proving their convergence rate if possible.

The method is widespread in the literatures, and the ideas behind it are tremendously useful. Güler [7] used the method to design an accelerated proximal point method, which inspired and served as the foundation of Catalyst Acceleration for variance reduced method in machine learning. Nesterov [13] also used the technique to design an accelerated cubic regularized Newton's method. In (6.1.19) of Nesterov's book [14], it's also used to derive a method of accelerated mirror descent. And finally, Geovani N. et al [6] used the technique to derive an accelerated Newton's method for convex composite objective function.

■ Add Guler's 1992 new proximal point paper.
■ Cite it.

■ Add the paper: "Accelerating the cubic regularization of Newton's method on convex problems"
■ Cite it.

■ Cite it.

■ Add paper: "Accelerated regularized newton methods for minimizing composite convex functions"
■ Cite it.

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

Definition 2.19 (Nesterov's estimating sequence) *Let $\phi_k : \mathbb{R}^n \to \mathbb{R}$ for all $k \geq 0$ be a sequence of functions. We call this sequence of function a Nesterov's estimating sequence when it satisfies the 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 2.20 *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 $\forall k \geq 0$:

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

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

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

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 the property of Nesterov's estimating sequence 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

This section is really about stating the results of the draft paper and no proofs will be done here. Along with the content of the draft paper, we will also explain the origin and inspirations of the ideas.

This section is based on the theoretical aspects of our draft paper. It will introduce major results and claims achieved during our research in each of the subsections. All theorems and claims stated in this section have proofs in the draft paper. The proofs haven't been carefully verified by authoritative people other than the author yet. We will start introducing the context and ideas for our research next.

Assume we want to solve a convex optimization problem: $\min_{x\in\mathbb{R}^n}\{F(x)\}$ and $F:\mathbb{R}^n\to\mathbb{R}$ is a L Lipschitz smooth function. We made this assumption for now for a faster exposition. One of the prime candidate for solving the optimization problem is the Nesterov's Accelerated Gradient methods (NAG) finds extensions for nonsmooth function through the proximal gradient operator. Proposed back in 1983 the original Nesterov's acceleration method [12] which uses the previous iterates to extrapolate the next iterate to evaluate the gradient. It's well known that if minimizer x^* exists for F, the method achieves a $\mathcal{O}(1/k^2)$ convergence rate on the objective value $F(x_k)$. This convergence rate is considered optimal for all class of L Lipschitz smooth convex function [14]. The convergence rate guarantee is faster than $\mathcal{O}(1/k)$ exhibited by gradient descent.

- Add Nesterov's original paper.
- Cite it.

- Add Nesterov's new book.
- Cite it.

We cover the algorithm briefly. Initialize $x_1=y_1$ and $t_0=1$, the algorithm finds $(x_k)_{k\geq 1}$ for all $k\geq 1$ by:

$$x_{k+1}=y_k-L^{-1}\nabla F(y_k), \tag{3.1}$$

$$t_{k+1}=1/2\left(1+\sqrt{1+4t_k^2}\right), \tag{3.2}$$

$$\theta_{k+1}=(t_k-1)/t_{k+1}, \tag{3.3}$$

$$y_{k+1}=x_{k+1}+\theta_{k+1}(x_{k+1}-x_k). \tag{3.4}$$

Unfortunately, the algorithm sped up the convergence rate for all convex function, it becomes slower for the subset of $\mu>0$ strongly convex function. This drawback inspired a vast amount of literatures aims at improving, extending, and analyzing NAG. Restarting is a popular solution to address the issue of obtaining faster convergence rate when the objective function is strongly convex. Beck and Toubelle [4] mitigated the issue by restarting and showed that it still has a $\mathcal{O}(1/k^2)$ convergence rate, and it performs better empirically. See (5.2.2) Necoara et al. [11] and Aujol et al. [2] and references within for recent advancements in restarting accelerated proximal gradient algorithm.

- Add Beck 2009 FISTA original paper.
- Cite it.

- Add Necoara linear convergence
- Add Aujol 2024 Parameter free FISTA restart.
- Cite Section 5.2.2 for the former, cite the entirety for the latter.

Restarting the algorithm is not the entire picture. Let $F : \mathbb{R}^n \to \mathbb{R}$ be a L Lipschitz smooth and $\mu > 0$ function. As introduced previously, in Walkington's writing, he showed that there exists a variant of the Nesterov's accelerated gradient method that achieved a linear convergence rate of $\mathcal{O}((1 - \sqrt{\kappa})^k)$ where $\kappa = \mu/L$. This convergence rate is strictly better than $\mathcal{O}((1 - \mu/L)^k)$ for the method of gradient descent. However, this variant has a fixed momentum parameter $\theta_{k+1} = (\sqrt{\kappa} - 1)(\sqrt{\kappa} + 1)^{-1}$ back in Equation 3.1. The same variant also appears in Beck's book as V-FISTA [3], and Nesterov's book as (2.2.22) [14].

■ Add Neol J. Walkington's "Nesterov's Method for Convex Optimization".
■ Cite it.

■ Add Beck's first order textbook
☐ Cite 10.7.7.
■ Add Nesterov's textbook.
■ Cite it.

One final Mystery of the algorithm is the convergence of the iterates which also has much to do with the momentum sequence $(\theta_k)_{k\geq 0}$ displayed in Equation 3.1. Chambolle, Dossal [5] showed that by choosing sequence $(t_k)_{k\geq 1}$ to be $t_k = (n + a - 1)/a$ where $a > 2$ instead would give $(x_k)_{k\geq 0}$ weak convergence in Hilbert space. It's put as an open question on what happens to the iterates when $a = 2$.

■ Add paper: "On the convergence of the iterates of the..."
■ Cite them.

All of these seemingly raises a crucial question: "Is it possible to describe something about the NAG algorithm for a set of sequence that is non-traditional?"; rephrasing it int a more technical manner: "What is the weakest description of the momentum sequence (θ_k) such that we can still claim something of value about the NAG algorithm?"

3.1 Our Contributions, organizations

Our contributions are two folds, theoretical and practical. The results are based the assumption $F = f + g$ where $g : R^n \to \overline{\mathbb{R}}$ is convex, and f is an L-Lipschitz smooth and $\mu \geq 0$ strongly convex function. We relax the traditional choice of the sequence θ_k in Equation 3.1 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, 3.15, 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 there exists $\rho_k > 1$ such that our R-WAPG reduces to a variant of FISTA proposed in Chambolle, Dossal [5], and we are able to show the same

■ Add "On the convergence of the iterates of..."
■ Cite them.

convergence rate in Theorem 3.18. When $\rho_k = 1, \mu = 0$, R-WAPG reduces perfectly to FISTA by in Beck [4]. If $\mu > 0, \rho_k = 1$, it reduces to the V-FISTA by Beck [3]. In Theorem 3.19, 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.

■ Add beck's original FISTA Paper.
■ Cite it.

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} 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.

Notations, and assumptions now follows. For all the subsection that follows, we let $F := f + g$ to take the same assumptions as in Section 2.2.1. Recall T_L, $\mathcal{G}_L$ denotes the proximal gradient operator and the gradient mapping operator. Additional notations are defined in the assumption below:

Assumption 3.1 Choose any integer $k \geq 0$. Given x_k, y_k, v_k, we define the following quantities

$$g_k := L(y_k - T_L y_k), \tag{3.5}$$

$$l_F(x; y_k) := F(T_L y_k) + \langle g_k, x - y_k \rangle + \frac{1}{2L}\|g_k\|^2, \tag{3.6}$$

$$\epsilon_k := F(x_k) - l_F(x_k; y_k), \tag{3.7}$$

Observe that by convexity of F, $\epsilon_k \geq 0$ for all $x_k, L > 0$. To see, use Theorem 2.17 and let $y = y_k, x = x_k$ which gives:

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

☐ Finish the organizations here after this section is finished. Still need to explain the numerical experiments section.

Organization now follows. Section 3.2 provides a stepwise description of the R-WAPG iterative algorithm along with an inequality crucial to proving the convergence rate later. Section 3.3 introduce the definition of an R-WAPG sequence, which constraints all possible parameters used permitted by the algorithm. The section also states the full R-WAPG algorithm and an upper bound on $F(x_k) - F(x^*)$. Here, x_k is generated by the R-WAPG algorithm and x^* is the minimizer. Section 3.4 bring forward three equivalent forms of the R-WAPG algorithm making it more comparable with other Accelerated Proximal Gradient

method appeared in the literatures. Section 3.5 gives characterizations of specific R-WAPG sequence that leads to convergence of the R-WAPG algorithm in terms of the optimality gap. The section also identifies specific instance of permissible R-WAPG sequences where it fits with the FISTA, V-FISTA, and the algorithm proposed by Chambolle Dossal [5].

■ Cite the Chambolle, Dossal 2015 paper here again.

3.2 Building Blocks of R-WAPG

Definitions:

- R-WAPG stepwise definition.
- R-WAPG stepwise convergence claim.

To do:

- Explain what is what.

Definition3.2 describes the procedures of generating the iterates (v_{k+1}, x_{k+1}) given any (v_k, x_k) and parameter $\alpha_k \in (0,1), \gamma_k > 0$. Proposition 3.3 gives an inequality instrumental to the convergence rate analysis, with the same assumption as the definition.

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:

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

Proposition 3.3 (Stepwise Lyapunov)
Fix any integer $k \in \mathbb{Z}$. Given any v_k, x_k and $\gamma_k > 0$, invoke Definition 3.2 to obtain $v_{k+1}, x_{k+1}, y_k, \hat{\gamma}_{k+1}$. Fix any arbitrary $R_k \in \mathbb{R}$. Define:

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

Then it has for all $x^* \in \mathbb{R}^n$ where $F^* = F(x^*)$, the inequality:

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

3.3 R-WAPG Sequence and R-WAPG algorithm

The R-WAPG Algorithm and convergence claim. Definitions:

- R-WAPG Sequence.
- R-WAPG algorithm
- Convergence of the R-WAPG algorithm.

Todo:

- Explain what is what.

The Definition 3.5 gives the definition of an iterative algorithm we called: Relaxed Weak Accelerated Proximal Gradient (R-WAPG) algorithm which generates sequence $(x_k, v_k)_{k\geq 1}$ using the R-WAPG sequences $(\alpha_k)_{k\geq 0}, (\rho_k)_{k\geq 0}$ given in Defintion 3.4. Proposition 3.6 shows the upper bound of the optimality gap $F(x_k) - F(x^*)$ with (x_k) generated by the R-WAPG algorithm.

Definition 3.4 (R-WAPG sequences)
Assume $0 \leq \mu < L$. The sequences $(\alpha_k)_{k\geq 0}, (\rho_k)_{k\geq 1}$ are sequences parameterized by μ, L. They are valid for R-WAPG if all the following holds:

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

*We call $(\alpha_k)_{k\geq 0}, (\rho_k)_{k\geq 0}$ the **R-WAPG Sequences**.*

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 $(\alpha_k)_{k\geq 0}, (\rho_k)_{k\geq 0}$ be given by Definition 3.4. The algorithm generates a sequence of vector $(y_k, x_{k+1}, v_{k+1})_{k\geq 1}$ for $k \geq 1$ by the procedures:

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

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

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 $k = 1, 2, \cdots, N$, 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 $k = 1, 2, \cdots, N$:

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

3.4 Equivalent forms of R-WAPG algorithm

Definitions:

- R-WAPG Intermediate form.
- R-WAPG Similar triangle form.
- R-WAPG Momentum form.

Theorems:

- R-WAPG First equivalent form.
- R-WAPG Second equivalent form.
- R-WAPG Third equivalent form.

Lemmas Todo:

(i) Explain what is what.

Definitions 3.7, 3.9 and 3.11 are three equivalent representations of the R-WAPG algorithms. Propositions 3.13, 3.14 and 3.15 states the equivalences between the forms and the sufficient conditions for initial conditions of x_1, v_1 such equivalence holds. The remarks of the definitions identifies specific instances in the literatures where the Accelerated Proximal Gradient method were presented under this specific form.

- Add and explain this part.

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 $k = 1, 2, \cdots$

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

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 [17], 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.

■ Add them
■ Cite them.

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 $k = 1, 2, \cdots$

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

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 [14], but without the necessary graphical illustrations to clarify it. Finally, a similar triangle for formulation of FISTA can be found in Equation (2), (3), (4) in [5]. 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. [8] and Ahn and Sra [1].

■ Cite Chambolle Dossal 2015 again.

■ Add these. ☐ Cite these.

The term "Method of Similar Triangle" specific to this particular representation of AGM. For more information, see:

(i) 3.1, 4.1 in Jongmin Lee et al., "A geometric structure of Acceleration and its role in making gradients small"

(ii) The entirety of Ahn and Sra's paper "PPM interpretation of Nesterov's accelerated gradient. ".

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 $k = 1, 2, \cdots$

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

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 [14], however, the sequence $(\alpha_k)_{k\geq 0}$ would be given by a different rule. See Theorem 3.18 and Lemma 3.16 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.

Proposition 3.13 (First equivalent representation of R-WAPG) *If the sequence $(y_k, v_k, x_k)_{k\geq 1}$ is produced by R-WAPG (Definition 3.5), then the iterates can be expressed without $(\gamma_k)_{k\geq 1}, (\hat{\gamma}_k)_{k\geq 2}$, and for all $k \geq 1$ they are algebraically equivalent to*

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

Proposition 3.14 (Second equivalent representation of R-WAPG) *Let iterates $(y_k, x_k, v_k)_{k\geq 1}$ and sequence $(\alpha_k, \rho_k)_{k\geq 0}$ be given by Definition 3.7. Then for all $k \geq 1$, iterate y_k, x_{k+1}, v_{k+1} satisfy:*

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

Proposition 3.15 (Third equivalent representation of R-WAPG) *Let sequence $(\alpha_k, \rho_k)_{k\geq 0}$ and iterates $(x_k, v_k, y_k)_{k\geq 1}$ given by R-WAPG intermediate form*

(Definition 3.9). Then for all $k \geq 1$, the iterates $(x_{k+1}, y_{k+1})_{k\geq 1}$ are algebraically equivalent to:

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

If in addition, $v_1 = x_1$ then

$$y_1 = \left(1 + \frac{L - L\alpha_1}{L\alpha_1 - \mu}\right)^{-1} \left(v_1 + \left(\frac{L - L\alpha_1}{L\alpha_1 - \mu}\right) x_1\right) = x_1.$$

In the special case when $\mu = 0$, the momentum term admits simpler representation

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

3.5 The descriptive power of R-WAPG on existing variants

Lemmas:

(i) Inverted FISTA sequence is a R-WAPG sequence.

(ii) Constant R-WAPG sequence.

Theorems:

(i) Convergence with constant momentum.

(ii) Convergence with Chambolle, Dossal Sequences.

Todo:

(i) Explain what is what.

Lemma 3.16 identifies a sequence $(\alpha_k)_{k\geq 0}$ such that $\alpha_k^{-2} \geq \alpha_{k+1}^{-2} - \alpha_{k+1}^{-1}$ as a specific instance of R-WAPG sequence. The lemma showed that sequence $(\alpha_k^{-1})_{k\geq 0}$ is the FISTA sequence which governs the momentum term and convergence claim in FISTA algorithms and variants alike. The lemma also provides a simplified convergence claim using the R-WAPG sequence on Proposition 3.6. Theorem 3.18 stated that the sequences given in Chambolle, Dossal's paper [5] indeed is an instance of R-WAPG sequence, along with that, it indeed attains a convergence rate $\mathcal{O}(1/k^2)$.

■ Cite this paper again here.

Lemma 3.16 (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) $\prod_{i=1}^{k} \max(1, \rho_{k-1})(1 - \alpha_k) = \alpha_k^2 \quad (\forall k \geq 1)$.

Lemma 3.17 (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) *Fix any* $r \in \left(\sqrt{q}, \sqrt{q^{-1}}\right)$ *then 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 term in Definition 3.11, which we denoted by* θ *has:* $\theta = (1 - r^{-1}\sqrt{q})(1 - r\sqrt{q})(1 - q)^{-1}$.

(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 3.18 (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 $k = 1, 2, \cdots$, *update:*

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

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 $\mathcal{O}(\alpha_k^2)$.

Theorem 3.19 (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 3.17. 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 $k = 1, 2, \cdots, N$, update:

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

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

4 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.

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.

Algorithm 1 Free R-WAPG

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

4.1 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 3.15 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.

4.1.1 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: $A = \text{diag}(0, \mu + (L-\mu)(N-1)^{-1}, \mu + 2(L-\mu)(N-1)^{-1}, \cdots, \mu + (N-2)(L-\mu)^{-1}, L)$.

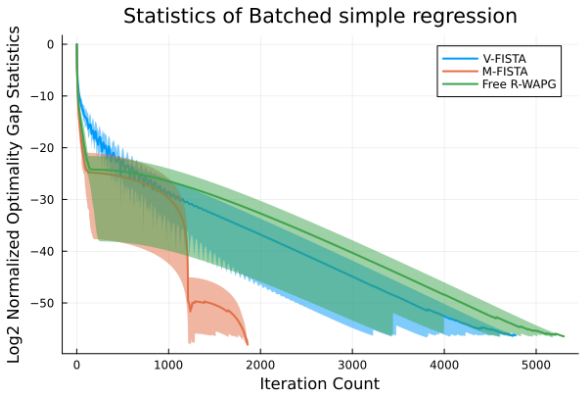
(iv) In this case $f = F = (1/2)\langle x, Ax\rangle$ and $g \equiv 0$.

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

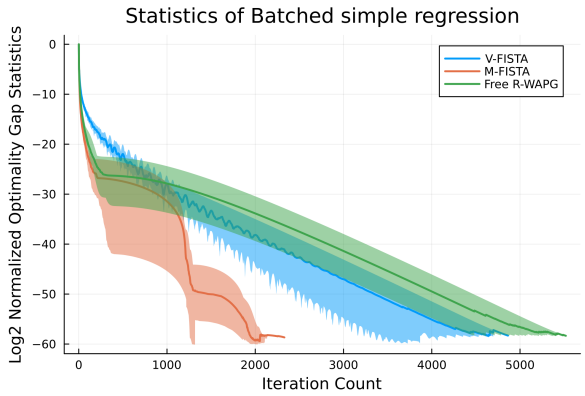
(vi) $x_0 \sim \mathcal{N}(I, \mathbf{0})$ is a vector, and it's the initial condition for all the algorithm. In this case the initial guess is fixed for all R-WAPG, M-FISTA and M-FISTA, but it's randomly generated by the zero mean standard normal distribution for each element in the vector.

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.

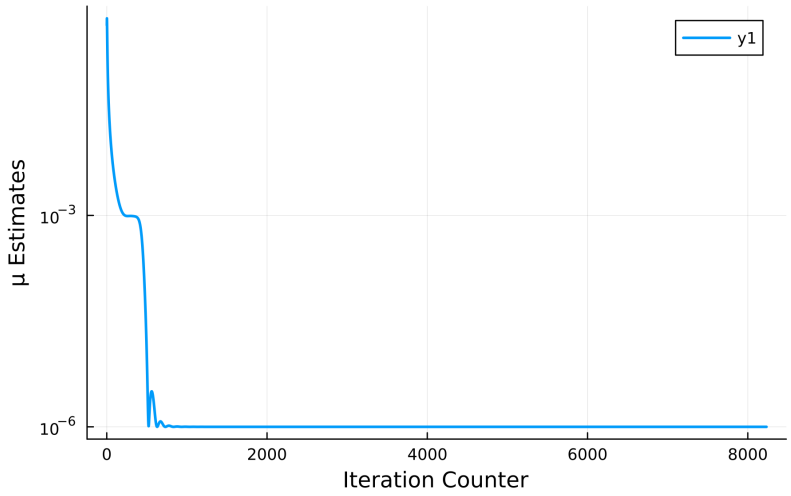


Figure 2: $N = 1024$, the μ estimates produced by Algorithm 1 (R-WAPG) is recorded.

4.1.2 LASSO

This section present results of numerical experiment for solving the (least absolute shrinkage and selection operator) LASSO problem proposed by Tibshirani [18]. The problem of LASSO has smooth, nonsmooth additive and the problem is given by:

■ Add reference "Regression shrinkage and selection via the Lasso"
■ Cite it.

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

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 $L = \|A^TA\|^2$.

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

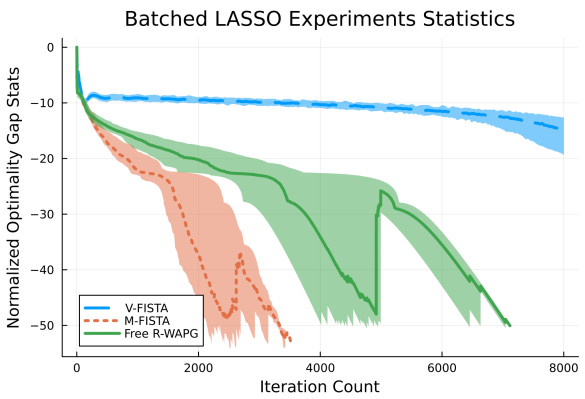
(v) Given x^+, it has $b = Ax^+ \in \mathbb{R}^M$.

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

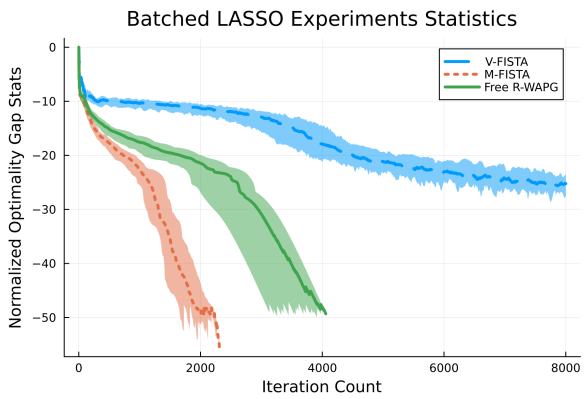
(vii) $x_0 \in \mathbb{R}^N$ is the initial guess. Its elements are random i.i.d variable realized from the standard normal distribution.

(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.



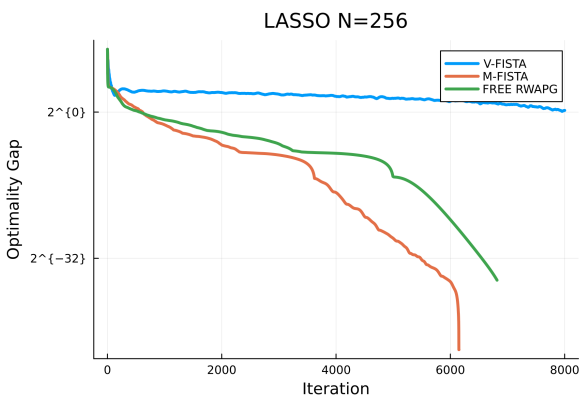
(a) LASSO experiment with $M = 64, N = 256$. Plots of minimum, maximum, and median δ_k with estimated F^*.



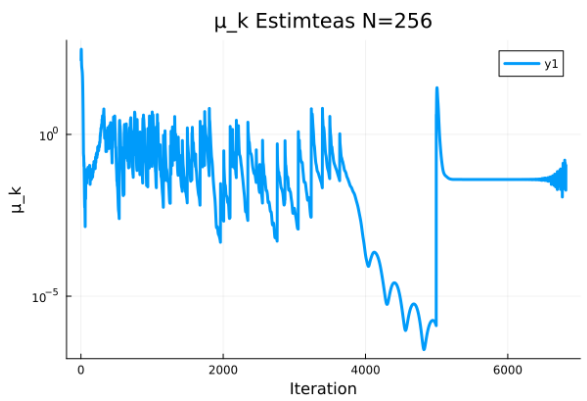
(b) LASSO experiment with $M = 64, N = 128$. Plots of minimum, maximum, and median δ_k with estimated F^*.

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.

4.1.3 Logistic regression

5 Catalyst accelerations and future works

Literatures review of the topics in Catalyst acceleration method. Here is a list of topics:

(i) The original accelerated PPM.

(ii) The Catalyst with weakly convex objectives.

After the literature reviews of the core literatures, move on and state new research directions and open problems. There are several directions for open problem:

(i) APPM method for monotone operators instead of just subgradient, whether the same framework exists in a greater context.

(ii) Accelerated Proximal Bregman Method.

(iii) Removing smoothness assumption in Catalyst acceleration framework.

A list of relevant literatures:

(i) Güler's 1992 paper on Accelerated Proximal Point method.

(ii) Lin's, and Payquette's three trilogy paper on Catalyst acceleration for convex, non-convex Variance reduced algorithm.

In this section, we assume that $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a convex function. This section reports major results from literatures concerning the method of Catalyst Acceleration.

Assumption 5.1 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. Define the model function for all $y \in \mathbb{R}^n$ to be

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

We define the Moreau Envelope at $y \in \mathbb{R}^n$ to be $\mathcal{M}^*_{\beta^{-1}}(y) = \min_{x\in\mathbb{R}^n} \mathcal{M}^{\beta^{-1}}(x; y)$. We denote $\mathcal{J}_{\beta^{-1}}$ to be the resolvent operator for subgradient of F, and $\mathcal{J}^{\epsilon}_{\beta^{-1}}$ for any $\epsilon \geq 0$ to be the inexact resolvent operator.

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

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

Inspired by accelerated proximal point method from Güler [7], and inexact proximal point method of Rockafellar 1976 [16], Lin [9] 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 for more information on variance reduced methods in machine learning. In brief, a variance reduce method (VRM) is a type of stochastic gradient methods that stablizes the current estimate using information of the gradient at previous iterates. In each iteration, only the gradient of a few samples are accessed but attain overall better complexity because the variance of the estimated gradients are reduced.

■ Add the bib for Mark Schmidt introductory paper RV methods.
☐ Cite it.

We can then list some citations and references to commonly known variance reduced method here.

The parts coming will introduce the algorithms, explain key innovations behind the algorithm, and key variants of the algorithm. The section at the end will give future works and extensions of the Catalyst framework.

Definition 5.2 (Lin's Universal Catalyst Acceleration)
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:

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

Remark 5.3 The above algorithm is Algorithm 1 from Lin's first paper on Catalyst Acceleration [9]. 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 $\mathcal{J}_{\kappa^{-1}}^{\epsilon_k} y_{k-1}$. 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 $\mu = 0$.

Much of the heavy lifting of constructing the algorithms were done through the method of estimating sequence introduced back in Definition 2.19. The convergence of the algorithm made use of an inexact version of the proximal gradient inequality (similar to Theorem 2.17) stated as Lemma A.7 in [9]. This lemma is instrumental for deriving an inexact variant of the estimating sequence $\phi_k^* \geq F(x_k) + \xi_k$. The sequence ϵ_k is compacted and bounded under

an upper bound parameterized via the sequence $(\epsilon_k)_{k\geq 0}$. The convergence proof from Lin was inspired by Schmidt's Inexact Proximal Gradient method.

■ Add bib, the paper is
□ Cite.

The technique of Estimating sequence results in depressingly long proof making it unsuitable for exposition here, however every significant piece of innovation is covered in details in our most recent Fall Winter 2024 MATH 590 report.

The following parts will complement the report and describe ideas on Lin's idea for choosing, termination criteria of the inner loop to evaluate $\mathcal{J}_k^{\epsilon_k} y_k$ in his second paper on Catalyst Acceleration [10]. To elucidate the matters consider the model function $\mathcal{M}^{\kappa^{-1}}(x;y)$ is $\mu+\kappa$ strongly convex, and therefore it admits the Polyak error bound condition:

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

By choosing x such that $\mathrm{dist}(\mathbf{0}, \partial\mathcal{M}^{\kappa^{-1}}(x;y)) \leq \epsilon$, it ensures $x \in \mathcal{J}^{\epsilon}_{\kappa^{-1}}(y)$. Unfortunately in practice, this is not used because a full gradient evaluation on the model function $\mathcal{M}^{\kappa}(\cdot;y)$ is costly (compare to the small amount required for variance reduced incremental method), so Lin suggested alternatives of Inner Loop Termination Criteria to make Catalyst acceleration competitive in practice.

Elucidate things by considering:

(i) What is the suggested error sequence ϵ_k for different type of objective function?

(ii) What are some conditions for achieving inexact evaluation up to the given accuracy ϵ_k?

(iii) What are the convergence claims for these different types of termination criteria.

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] 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).

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

[4] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 183–202.

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

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

[8] 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.

[9] 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.

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

[11] 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.

[12] Y. Nesterov, *A method for solving the convex programming problem with convergence rate O(1/k^2)*, Proceedings of the USSR Academy of Sciences, (1983).

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

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

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

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

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

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

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