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

Hongda Li

Department of Mathematics
University of British Columbia,
Okanagan Campus.

December 28, 2024

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 [2], 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, 4 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 4 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 citations here.

Section 6, 7, 8 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.

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

2.1 Fundamentals in non-convex analysis

We are in $\mathbb{R}^n$, and the weakest assumption we are making for the objective function is Local Lipschitz continuity. Definitions:

- ■ Local Lipschitz continuity.
- □ Regular subgradient. Remember to cite.
- □ Limiting subgradient. Remember to cite.
- ■ Weakly convex function.
- ■ The Bregman Divergence of function.

Take Limiting, Regular subgradient definitions from Cui, Pong's book, Definition 4.3.1.

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*

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

Definition 2.3 (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*

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

Definition 2.4 (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 2.5 (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:*

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

2.2 Fundamentals in convex analysis

Introduce

- Convexity,
- convex subgradient,
- Lipschitz smoothness.

Definitions:

- Strong convexity of a function.
- The proximal gradient operator.
- The proximal mapping operator.

Lemmas:

- Quadratic growth conditions of a strongly convex function.

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 $\mathrm{ri}(\mathrm{dom}(F))$. So it has $\mathrm{ri}(\mathrm{dom}\, F) \subseteq \mathrm{dom}(\partial F) \subseteq \mathrm{dom}\, F$.

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 2.6 (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 $y \in \mathrm{dom}(\partial F)$, 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 2.7 (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 2.8 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

Introduce notations for the proximal gradient model function. Lemmas:

- Proximal gradient envelope.
- A property of gradient mapping.

Theorems:

- The proximal gradient inequality.

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 2.9 (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 2.10 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 2.11 (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 2.12 (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 2.13 (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 2.14 This lemma still holds for non-convex f under prox-boundness and weak convexity and differentiability of f.

Lemma 2.15 (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 2.16 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 counter parts assumes a convex function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$. Further discussions would require giving an example of estimating sequence first. The definition follows gives an example of an estimating sequence.

Definition 2.17 (Nesterov's estimating sequence)

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 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 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

Cite Nesterov's original paper on this.

Cite Chapter 2 of Nesterov's new book.

L Lipschitz smooth convex function. The convergence rate guarantee is faster than $\mathcal{O}(1/k)$ exhibited by gradient descent.

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 mitigate the issue by restarting and showed that it still has a $\mathcal{O}(1/k^2)$ convergence rate, and it performs better empirically. See and references within for recent advancements in restarting accelerated proximal gradient algorithm.

Cite Beck 2009 FISTA.

Cite Necoara linear convergence, and Aujol 2024 Parameter free FISTA restart.

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, and Nesterov's book as (2.2.22).

Cite Walkington's education stuff here.

Cite them.

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 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 $a = 2$.

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, 3.12, 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 [?], and we are able to show the same convergence rate in Theorem 3.15. When $\rho_k = 1, \mu = 0$, R-WAPG reduces perfectly to FISTA by Beck [1], if $\mu > 0, \rho_k = 1$, it reduces to the V-FISTA by Beck [1]. In Theorem 3.16, 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.

Fix citation here.

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 quan-

tities

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

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

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

Observe that by convexity of F, $\epsilon_k \geq 0$ for all $x_k, L > 0$. To see, use Theorem 2.15 and let $y = y_k, x = x_k$ which gives:

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

Organization now follows.

Finish the organizations here after this section is finished.

3.2 Building Blocks of R-WAPG

Definitions:

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

Todo:

(i) Explain what is what.

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:

(i) Explain what is what.

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.

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 $k = 1, 2, \cdots$

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

Definition 3.8 (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 $k = 1, 2, \cdots$

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

Definition 3.9 (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 $k = 1, 2, \cdots$

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

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

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

Proposition 3.10 (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

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

Proposition 3.11 (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:

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

Proposition 3.12 (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 3.8). 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 3.13 (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 3.14 (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 3.9, 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 3.15 (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 3.16 (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 3.14. 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 Method Free R-WAPG

Algorithm, and results of numerical experiments with their descriptions.

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 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 triology paper on Catalyst acceleration for convex, non-convex Variance reduced algorithm.

6 Performance estimation problems

There are several foundational papers relevant.

7 Methods of inexact proximal point

8 Nestrov's acceleration in the non-convex case

9 Using PostGreSQL and big data analytic method for species classification on Sentinel-2 Satellite remote sensing imagery

References

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

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