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

Hongda Li

Department of Mathematics
University of British Columbia,
Okanagan Campus.

December 22, 2024

## Contents

# 1 Introduction

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

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

Unless specified, assume $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is $L$-Lipscthiz smooth $\mu \geq 0$ strongly convex and $g : Q \rightarrow \overline{\mathbb{R}}$ 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 [1], 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 $L, \mu$ 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 specify 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” occuring 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 $F : O \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ 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 $F : O \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ 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 $x, y \in \mathcal{N}(\bar{x})$: $|F(x) - F(y)| \le K\|x - y\|$.*

**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*

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

**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*

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

**Definition 2.4 (Weakly convex function)** *$F : \mathbb{R}^n \to \overline{\mathbb{R}}$ is $\mu$ 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 convex in this section. When $F$ is convex, the limiting subgradient and the regular subgradient reduced to the following definition:

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

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:

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

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

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

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 $\mu \ge 0$ strongly convex if and only if for any fixed $y \in \operatorname{dom}(\partial F)$, we have for all $x \in \mathbb{R}^n$:*

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

**Lemma 2.7 (Quadratic growth from strong convexity)** *If $F$ is $\mu \ge 0$ strongly convex, $\bar{x}$ is a minimizer of $F$. Then for all $x \in \mathbb{R}^n$*

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

**Remark 2.8** The minimizer is unique whenever $\mu > 0$. For contradiction, assume $x$ is another minimizer, then $F(x) \ne F(\bar{x})$, which is a direct contradiction.

### 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 $\mu \ge 0$ strongly convex and $g : \mathbb{R} \to \overline{\mathbb{R}}$ is convex. To make the discussion simpler, fix any $\beta \ge 0$ we define the following model functions as a $\mathbb{R}^n \times \mathbb{R}^n \to \overline{\mathbb{R}}$:

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

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)** *Suppose that $F := f + g$ where $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex and $f : \mathbb{R}^n \to \mathbb{R}$ is a $L$ Lipschitz smooth function. Define the proximal gradient operator $T_L$ on all $y \in \mathbb{R}^n$:*

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

**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 $f$.

**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.

# 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.

Definitions:

(i) Method of Nesterov's estimating sequence.

(ii) R-WAPG Sequence.

(iii) R-WAPG algorithm.

(iv) R-WAPG Intermediate form.

(v) R-WAPG Similar triangle form.

(vi) R-WAPG Momentum form.

Theorems:

(i) Convergence of the R-WAPG algorithm.

(ii) R-WAPG First equivalent form.

(iii) R-WAPG Second equivalent form.

(iv) R-WAPG Third equivalent form.

(v) Convergence with constant momentum.

(vi) Convergence with Chambolle, Dossal Sequences.

Lemmas

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

(ii) Constant R-WAPG sequence.

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. 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 $F:\mathbb{R}^n\rightarrow\mathbb{R}$ 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 $L$ Lipschitz smooth convex function. The convergence rate gurantee is faster than $\mathcal{O}(1/k)$ exibited by gradient descent.

Cite Nesterov's original paper on this.

Cite Chapter 2 of Nesterov's new book.

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 empirically better. 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 $\mathcal{O}((1 - \sqrt{\mu/L})^k)$. However, this variant has a fixed momentum parameter $\theta_{k+1} = \frac{\sqrt{\kappa}-1}{\sqrt{k}+1}$ back in Equation 3. 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. Chambolle, Dossal showed that by choosing sequence $(t_k)_{k\geq 1}$ to be $t_k =$

### 3.1 Our Contributions

## 4 Method Free R-WAPG

Algorithm, and results of numerical experiments with their descriptions.

# 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.

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

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