First Order Nonsmooth Optimization: Catalyst Acceleration and Unifying Nesterov's Acceleration

Hongda Li

University of British Columbia Okanagan

January 24, 2025

Overview

This talk will be based on the content of our draft paper and selected content of the Catalyst Meta Acceleration Framework.

Our preprint:

1. X. Wang and H. Li, *A Parameter Free Accelerated Proximal Gradient Method Without Restarting*, preprint, (2025).

Catalyst Meta Acceleration:

1. H. Lin, J. Mairal and Z. Harchaoui, *A universal catalyst for first-order optimization*, in NISP, vol. 28, (2015).
2. ________, *Catalyst acceleration for first-order convex optimization: from theory to practice*, JMLR, 18 (2018), pp. 1–54.

ToC

Notations and preliminaries

Throughout this talk, let $\mathbb{R}^n$ be the ambient space equiped with Euclidean inner product and norm. We consider

$$\min_{x\in\mathbb{R}^n} \{F(x) := f(x) + g(x)\}\,. \tag{1}$$

Unless specified, assume:

1. $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipscthiz smooth $\mu \geq 0$ strongly convex,
2. $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is closed convex proper.

Notations and preliminaries

Definition 1 (Proximal gradient operator)

Define the proximal gradient operator T_L on all $y \in \mathbb{R}^n$:

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

Definition 2 (Gradient mapping operator)

Define the gradient mapping operator $\mathcal{G}_L$ on all $y \in \mathbb{R}^n$:

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

Proximal gradient inequality

Lemma 3 (The proximal gradient inequality)
For all $y \in \mathbb{R}^n$, $x \in \mathbb{R}^n$, it has:

$$F(x) - F(T_L y) - \langle L(y - T_L y), x - y \rangle - \frac{\mu}{2}\|x - y\|^2 - \frac{L}{2}\|y - T_L y\|^2 \geq 0.$$

This lemma is crucial to developing results in our current draft paper.

Nesterov's estimating sequence example

Definition 4 (Nesterov's estimating sequence)

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

1. There exists another sequence $(x_k)_{k\geq 0}$ such that for all $k \geq 0$ it has $F(x_k) \leq \phi_k^* := \min_x \phi_k(x)$.
2. 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 $\phi_{k+1}(x) - \phi_k(x) \leq -\alpha_k(\phi_k(x) - F(x))$.

The technique is widespread in the literatures and it's used to derive the convergence rate of acceleration on first order method, and the numerical algorithm itself. It is a two birds one stone technique.

Our works on R-WAPG

Here are the contributions of our draft paper. Recall the Nesterov's acceleration has momentum extrapolation updates on $y_{k+1} = x_{k+1} + \theta_{k+1}(x_{k+1} - x_k)$. We proposed the idea of R-WAPG, a generic method that:

1. Describe for momentum sequences that doesn't follow Nesterov's rules.
2. Unifies the convergence rate analysis for several Euclidean variants of the FISTA method.
3. A parameter free numerical algorithm: "Free R-WAPG" method that has competitive numerical performance in practical settings without restarting.

Our work is inspired by considering Nesterov's estimating sequence where $F(x_k) + R_k = \phi_k^*$.

Introducing Catalyst Part I

Introducing Catalyst

Let $F : \mathbb{R} \to \overline{\mathbb{R}}$ be $\mu \geq 0$ strongly convex and closed. Let the initial estimate be $x_0 \in \mathbb{R}^n$, fix parameters $\kappa > 0$ and $\alpha_0 \in (0, 1]$.

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

$$\text{find } x_k \approx \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ F(x) + (\kappa/2)\|x - y_{k-1}\|^2 \right\},$$

$$\text{find } \alpha_k \in (0,1) \text{ such that } \alpha_k^2 = (1 - \alpha_k)\alpha_{k-1}^2 + (\mu/(\mu+\kappa))\alpha_k,$$

$$y_k = x_k + \frac{\alpha_{k-1}(1 - \alpha_{k-1})}{\alpha_{k-1}^2 + \alpha_k}(x_k - x_{k-1}).$$

We will return to this in the later slides.

Introducing Catalyst Part II

Catalyst by Lin, et al. [6, 5] has the theoretical and pratical importance:

1. It's an early attempt at putting accelerated inexact proximal point method into a practical settings.
2. It finds application in machine learning and it accelerates the convergence of Varianced Reduced Method (A type of incremental method that is not slower than the exact counter part).
3. It demonstrates crucial ideas on how prove convergence rate where the evaluation of proximal point method is inexact in the convex settings.

R-WAPG sequences

Definition 5 (R-WAPG sequences)

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

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

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

The method of R-WAPG

Definition 6 (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 5. 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 $k = 1, 2, 3, \ldots$

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

Convergence of R-WAPG

The convergence claim of the method follows.

Proposition 2.1 (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 6. Define $R_1 = 0$ and suppose that for $k = 1, 2, \ldots, N$, we have R_k recursively given by:

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

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

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

Equivalent forms of R-WAPG

1. Equivalent forms of R-WAPG exists and resembles variants of FISTA in the literatures
2. We proved the equivalences in our draft papers and the convergence claim from previous applies to all the equivalent forms of R-WAPG which will follow.

R-WAPG intermediate form

Definition 7 (R-WAPG intermediate form)

Assume $\mu < L$ and let $(\alpha_k)_{k\geq 0}, (\rho_k)_{k\geq 0}$ given by Definition 5. Initialize any x_1, v_1 in $\mathbb{R}^n$. For $k \geq 1$, the algorithm generates sequence of vector iterates $(y_k, v_{k+1}, x_{k+1})_{k\geq 1}$ by the procedures:

For $k = 1, 2, \ldots$

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

R-WAPG intermediate form

Definition 7 (R-WAPG intermediate form)

Assume $\mu < L$ and let $(\alpha_k)_{k\geq 0}, (\rho_k)_{k\geq 0}$ given by Definition 5. Initialize any x_1, v_1 in $\mathbb{R}^n$. For $k \geq 1$, the algorithm generates sequence of vector iterates $(y_k, v_{k+1}, x_{k+1})_{k\geq 1}$ by the procedures:

For $k = 1, 2, \ldots$

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

1. If, $\mu = 0$, this is Chapter 12 of in Ryu and Yin's Book [7], right after Theorem 17.

R-WAPG similar triangle form

Definition 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 5. For $k \geq 1$, the algorithm generates sequences of vector iterates $(y_k, v_{k+1}, x_{k+1})_{k\geq 1}$ by the procedures:

For $k = 1, 2, \ldots$

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

R-WAPG similar triangle form

Definition 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 5. For $k \geq 1$, the algorithm generates sequences of vector iterates $(y_k, v_{k+1}, x_{k+1})_{k\geq 1}$ by the procedures:

For $k = 1, 2, \ldots$

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

1. Equation (2), (3), (4) in [3] is a similar triangle formulation of FISTA with $\mu = 0$.
2. see (3.1, 4.1) in Lee et al. [4] and Ahn and Sra [1] for graphical visualization of similar triangle form.

R-WAPG momentum form

Definition 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 5. The algorithm generates iterates x_{k+1}, y_{k+1} For $k = 1, 2, \cdots$ by the procedures:

For $k = 1, 2, \ldots$

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

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

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

Summary of our results

With the equivalent representations and the convergence claim for relaxed sequence $(\alpha_k)_{k\geq 0}$ of the R-WAPG, we are able to unifies:

1. Several Euclidean variants of the FISTA algorithm.
2. Nontraditional choices of momentum sequences.

The table below summarizes our major results.

Algorithm	μ	α_k, ρ_k	$F(x_k) - F^* \leq \mathcal{O}(\cdot)$
Definition 6	$\mu \geq 0$	$\alpha_k \in (\mu/L, 1)$, $\rho_k > 0$	$\prod_{i=0}^{k-1} \max(1, \rho_i)(1 - \alpha_{i+1})$ (Proposition 2.1)
FISTA [3]	$\mu = 0$	$0 < \alpha_k^{-2} \leq \alpha_{k+1}^{-1} - \alpha_{k+1}^{-2}$, $\rho_k \geq 1$	α_k^2
V-FISTA (10.7.7) [2]	$\mu > 0$	$\alpha_k = \sqrt{\mu/L}$, $\rho_k = 1$	$(1 - \sqrt{\mu/L})^k$,
Definition 6	$\mu > 0$	$\alpha_k = \alpha \in (\mu/L, 1)$, $\rho_k = \rho > 0$	$\max(1 - \alpha, 1 - \mu/(\alpha L))^k$

These results are consistent of iteratures. To the best of our knowledge, the last variant is a and we have the convergence claim for it using R-WAPG.

Free R-WAPG

Citation examples

Citation examples [3]

References I

- K. Ahn and S. Sra, *Understanding Nesterov's acceleration via proximal point method*, in Symposium on Simplicity in Algorithms, SIAM, June 2022, pp. 117–130.
- A. Beck, *First-order Methods in Optimization*, MOS-SIAM Series in Optimization, SIAM, 2017.
- 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.
- 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, 2021, pp. 11999–12012.

References II

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.

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

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