Nesterov's First Order Method Accelerations: Unifications, Applications and Numerial Experiments

Hongda Li

University of British Columbia Okanagan

February 9, 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 I

Introduction

Part I: Introduction.

Upcoming Contents:

- ▶ Notations, proximal gradient inequality, gradient mapping.
- ▶ Nesterov's estimating sequence.
- ▶ Introducing R-WAPG (Content of our draft paper).
- ▶ Introducing the Catalyst metal acceleration framework.

Notations and preliminaries

Throughout this talk, let $\mathbb{R}^n$ be the ambient space equipped 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.

3. Minimum exists $F^* = \min_{x \in \mathbb{R}^n} \{f(x) + g(x)\}$ and minimizer exists.

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

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:

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

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

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

We will return to this in the later slides.

Introducing Catalyst Part II

Catalyst by Lin, et al. [10, 9] has the theoretical and practical importance:

1. It puts accelerated inexact proximal point method into a practical setting.

2. It finds application in machine learning, and it accelerates the convergence of Variance Reduced Method (A type of incremental method that is not slower than the exact counterpart).

3. It demonstrates crucial ideas on convergence analysis involving nexact evaluation of proximal point method under the the convex settings.

Contents of our draft papers

Part II: Contents of our current draft paper.

Upcoming contents

We explain results from our relaxed, weak accelerated proximal gradient (R-WAPG).

1. The R-WAPG sequence and the method of R-WAPG.
2. A generic convergence theorem of R-WAPG.
3. Three equivalent forms of R-WAPG.
4. Specialized convergence results with for specific instances of FISTA variants in the literatures.
5. A parameter free formulation of R-WAPG and the results of our preliminary numerical experiments.
6. Nesterov's idea of strong convexity transfer and the R-WAPG framework.

R-WAPG sequences

Definition 5 (R-WAPG sequences)

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

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

We call $(\alpha_k)_{k \geq 0}, (\rho_k)_{k \geq 0}$ 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 $k = 1, 2, 3, \ldots$

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

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

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

Equivalent forms of R-WAPG

1. Equivalent forms of R-WAPG exists and resembles variants of FISTA in the literatures
2. We proved the upcoming equivalences forms of R-WAPG in the draft papers.
3. Previous claimed generic convergence results will apply to all equivalent forms of R-WAPG will now come.

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 $k = 1, 2, \ldots$

$$y_k = \left(1 + \frac{L - L\alpha_k}{L\alpha_k - \mu}\right)^{-1} \left(v_{k+1} + \left(\frac{L - L\alpha_k}{L\alpha_k - \mu}\right) x_k\right),$$

$$x_{k+1} = y_k - L^{-1}\mathcal{G}_L y_k,$$

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

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 $k = 1, 2, \ldots$

$$y_k = \left(1 + \frac{L - L\alpha_k}{L\alpha_k - \mu}\right)^{-1} \left(v_{k+1} + \left(\frac{L - L\alpha_k}{L\alpha_k - \mu}\right) x_k\right),$$

$$x_{k+1} = y_k - L^{-1} \mathcal{G}_L y_k,$$

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

1. If, $\mu = 0$, this is Chapter 12 of in Ryu and Yin's Book [15], 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 $k = 1, 2, \ldots$

$$y_k = \left(1 + \frac{L - L\alpha_k}{L\alpha_k - \mu}\right)^{-1} \left(v_k + \left(\frac{L - L\alpha_k}{L\alpha_k - \mu}\right) x_k\right),$$

$$x_{k+1} = y_k - L^{-1}\mathcal{G}_L y_k,$$

$$v_{k+1} = x_{k+1} + (\alpha_k^{-1} - 1)(x_{k+1} - x_k).$$

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 $k = 1, 2, \ldots$

$$y_k = \left(1 + \frac{L - L\alpha_k}{L\alpha_k - \mu}\right)^{-1} \left(v_k + \left(\frac{L - L\alpha_k}{L\alpha_k - \mu}\right) x_k\right),$$

$$x_{k+1} = y_k - L^{-1} \mathcal{G}_L y_k,$$

$$v_{k+1} = x_{k+1} + (\alpha_k^{-1} - 1)(x_{k+1} - x_k).$$

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. [8] 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
$k = 1, 2, \ldots$ by the procedures:

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

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

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

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

1. Three Euclidean variants of the FISTA algorithm.
2. Constant momentum APG for strongly convex objective.

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 literatures. To the best of our knowledge, the last variant is, and we have the convergence claim for it using R-WAPG.

Free R-WAPG

We proposed the following implementation of R-WAPG which doesn't require parameters μ, L in advance.

Algorithm 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 $L/2\|x^+ - y_k\|^2 < D_f(x^+, y_k)$ **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

Simple quadratic optimizations

Our metric is called normalized optimality gap:

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

This is our first numerical experiment is:

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

Our setup has:

1. Same initial guess x_0 shared by FR-WAPG, M-FISTA and V-FISTA in (10.7.7, 10.7.6) by Beck [2] repeated with 30 different random initial guesses.

2. Min, max, and median of δ_k measured for all iterative method.

3. μ, L were given in prior to produce diagonal matrix

$$A = \text{diag}\left(0, \mu + \frac{L - \mu}{N - 1}, \mu + 2\frac{L - \mu}{N - 1}, \ldots, \mu + \frac{(N - 2)(L - \mu)}{N - 1}, L\right).$$

M-FISTA, FR-WAPG were not fed these parameters.

Simple quadratic optimizations results

We had $L = 1, \mu = 10^{-5}$ and this is the results:
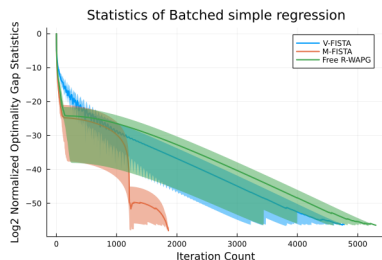
(a) $N = 256$, simple convex quadratic.
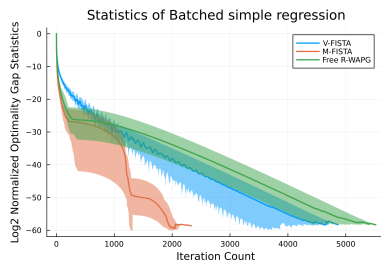
(b) $N = 1024$, simple convex quadratic.

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

FR-WAPG μ estimation graph

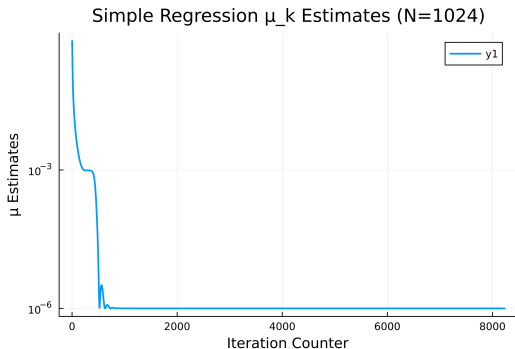FR-WAPG estimates the following for μ during its execution:

Simple Regression μ_k Estimates (N=1024)

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

LASSO numerical experiment

Tibshirani [17] proposed:

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

Our setup:

1. M, N are constants. $A \in \mathbb{R}^{M \times N}$ has i.i.d random entries from a standard normal distribution.

2. L, μ, are estimated by A by $\mu = 1/\|(A^T A)^{-1}\|$ and $L = \|A^T A\|$.

3. The synthetic solution is $x^+ = [1 \ -1 \ 1 \ \cdots]^T \in \mathbb{R}^N$ and $b = Ax^+ \in \mathbb{R}^M$.

4. $x_0 \in \mathbb{R}^N$ is the initial guess with i.i.d random variable from a standard normal distribution. Same initial guess shared by FR-WAPG, V-FISTA, M-FISTA.

LASSO numerical experiment results

Recorded statistics of δ_k for all algorithms.
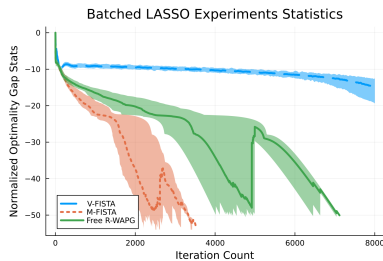
(a) LASSO experiment with $M = 64, N = 256$. Plots of minimum, maximum, and median δ_k with estimated F^*.
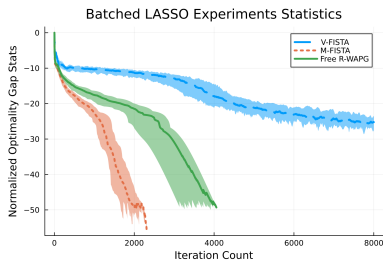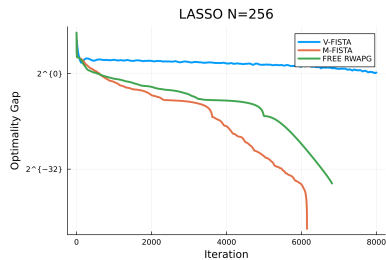
(b) LASSO experiment with $M = 64, N = 128$. Plots of minimum, maximum, and median δ_k with estimated F^*.
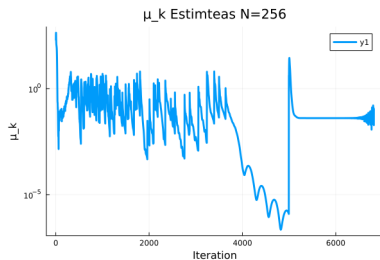
Figure: LASSO experiments.

μ estimates produced from a single LASSO experiment

FR-WAPG produces the following estimates for μ on one of the test instance of LASSO:

(a) Single lasso experiment plot of δ_k with.

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

Figure: A single LASSO experiment results, with $M = 64, 256$. We had $\mu = 7.432363627613958 \times 10^{-18}$ and $L = 2321.737206983643$.

Nesterov's idea of strong convexity transfer

There is one detail that our R-WAPG doesn't incorporate on all Euclidean variants of FISTA. We consider this a minor augmentation for the future.

1. In Nesterov's 2013 paper [12], he considers accelerated minimization problem $\phi = f + g$ with f L_f smooth and g being $\mu_g \geq 0$ strongly convex.

2. Algorithm 5, Chambolle, Pock [4] captures several variants of FISTA, and it assumes that $F = f + g$ where f, g has strong convexity constant $\mu_f \geq 0, \mu_g \geq 0$ respective so F is $\mu := \mu_f + \mu_g \geq 0$ strongly convex.

Fast linear convergence is possible if any one of f, g of the function is strongly convex, but this is not yet a prediction of R-WAPG.

Selected contents from Catalyst Meta Accelerations

Part III: Catalyst Acceleration and future works.

Upcoming Contents

1. Notations, assumptions.

2. Inexact proximal point evaluations criteria.

3. Structure of Catalyst accelerations and inner/out loop complexity results from Lin et al.'s first Catalyst paper [9].

4. Catalyst acceleration with relative error criterion, Lin et al.'s second paper [10].

5. Comments on future works for Catalyst by Necoara et al. [11] and our questions about the Catalyst Acceleration Framework.

Assumptions in Catalyst

Assumption 3.1

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

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

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

Absolute termination criterion

Definition 10 (Absolute termination criterion C1)

Take F as given by Assumption 3.1. For $\epsilon > 0, \kappa > 0$ and $x \in \mathbb{R}^n$, the absolute criterion C1 characterizes the set of inexact proximal iterates by:

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

Definition 11 (Relative termination criterion C2)

Take F as given by Assumption 3.1. Given any $\delta \in (0, 1]$, $\kappa > 0$ and $x \in \mathbb{R}^n$, the relative criterion C2 of the inexact resolvent is defined by:

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

Catalyst Meta Acceleration

Definition 12 (Lin's Universal Catalyst Acceleration)

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

find $x_k \in \mathcal{J}^{\epsilon_k}_{\kappa^{-1}F} y_{k-1}$,

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

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

Notations and Variance Reduced Methods

Catalyst notations

Denotes

1. the outer loop algorithm, i.e: Definition 12 by $\mathbb{A}$, it generates $(x_k, y_k)_{k \geq 1}$;

2. the inner loop algorithm by $\mathbb{M}$ that evaluates $x_k \in \mathcal{J}_{\kappa^{-1}}^{\epsilon_k} y_{k-1}$, it generates $(z_{k,t})_{t \geq 0}$.

The class of variance reduced methods are often used for $\mathbb{M}$. They are examples of incremental methods. Major examples of VRMs include SVRG by Xiao, Zhang [18], Finito by Defazio et al. [6], SAG by Schmidt et al. [16], and SAGA by [5].

Inner loop complexity assumption

The following assumption is crucial, and we will return to it.

Assumption 3.2 (Linear convergence of inner loop)

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

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

Outer loop complexity, $\mu > 0$

This is Theorem 3.1 in Lin et al. [9].

Theorem 13

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

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

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

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

Outer loop complexity, $\mu \geq 0$

This is Theorem 3.3 from Lin et al. [9].

Theorem 14

For $\mathbb{A}$ with regularization parameter $\kappa > 0$. Assume that F is convex but with strong convexity constant $\mu \geq 0$. Choose $\alpha_0 = (\sqrt{5} - 1)/2$ and the absolute error sequence

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

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

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

Inner loop complexity, absolute errors

The followings are Proposition 3.2, 3.4 from Lin et al. [9].

Proposition 3.1 (Inner loop complexity $\mu > 0$)

Under the same settings of Theorem 13, suppose that

1. $\mathbb{M}$ has linear convergence rate as specified in Assumption 3.2,
2. $\mathbb{M}$ is initialized with $z_{k,0} = x_{k-1}$ for all $k \geq 2$.

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

Proposition 3.2 (Inner loop complexity $\mu \geq 0$)

Under the settings of Theorem 14, suppose that:

1. $\mathbb{M}$ has linear convergence rate as specified in Assumption 3.2,
2. the initial guess for $\mathbb{M}$ is $z_{0,k} = x_{k-1}$ and F has bounded level set.

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

Total complexity

We count m, the number of iteration experienced by $\mathbb{M}$ up to the k iteration of $\mathbb{A}$.

1. If $\mu > 0$, 3.1 gives $m \leq T_{\mathbb{M}} k$. Substituting $k \geq m/T_{\mathbb{M}}$ into Theorem 13:

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

2. If $\mu = 0$ so F is not necessarily strongly convex, using Proposition 3.2, Theorem 14 it has

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

So:

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

These results are stated in Section 3.2 in Lin et al. [9].

Catalyst acceleration with relative errors

Definition 15 (Catalyst Acceleration with relative error)

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 $\alpha_0 = \sqrt{q}$ where $q = \mu/(\kappa + \mu)$. Let $(\delta_k)_{k \geq 0}$ be an absolute 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:
>
> find $x_k \in \widetilde{\mathcal{J}}_{\kappa^{-1}F}^{\delta_k} y_{k-1}$,
>
> find $\alpha_k \in (0, 1)$ such that $\alpha_k^2 = (1 - \alpha_k)\alpha_{k-1}^2 + q\alpha_k$,
>
> $y_k = x_k + \dfrac{\alpha_{k-1}(1 - \alpha_{k-1})}{\alpha_{k-1}^2 + \alpha_k}(x_k - x_{k-1})$.

Outer loop complexity under relative errors

Theorem 16 (Outer loop complexity under criterion C2)

For the iterates $(x_k)_{k \geq 0}$ generated by algorithm in Definition 15, we have

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

Remark 3.1

This is Proposition 8, 9 in Lin et al.'s second Catalyst paper [10]. For a precise description of the upper bound, see Theorem 8.

Observe that it doesn't require knowledge about F^*.

Inner loop warm start under criterion C2

Assumption 3.3 (Warm start conditions for the inner loop using C2)

For minimization sub problem $\mathcal{M}_F^{\kappa^{-1}}(z; y_k)$ solved by $\mathbb{M}$, we use the following warm start condition for $z_{k,0}$ under termination criterion C2 (Definition 11):

1. If F is smooth then choose $z_k = y_k$.
2. Else it's composite, define $f_k(x) = f(x) + \frac{\kappa}{2}\|x - y_k\|^2$ and do

$$z_{k,0} = \text{prox}_{\eta g}\left(y_k - \eta \nabla f_\kappa(y_k)\right) \quad \text{with } \eta = 1/(\kappa + L).$$

Remark 3.2

See pg 13 of Lin et al. for more detail about warm starting strategies of $\mathbb{M}$ under Termination criterion C1.

Inner loop complexity under C2

Theorem 17 (Inner loop complexity with relative error C2)

Consider $\mathbb{M}$ under the settings of Assumption 3.2 and Assumption 3.3, let the relative error sequence $(\delta_k)_{k \geq 0}$ be given by Theorem 16, and let $\mathbb{A}$ be defined as in Definition 15. Then the sequence $(z_{k,t})_{t \geq 0}$ generated by $\mathbb{M}$ has
$T_{k+1} = \min \left\{ t \geq 0 : z_{k,t} \in \widetilde{\mathcal{J}}_{\kappa^{-1}F}^{\delta} y_k \right\}$ *satisfies:*

$$T_{k+1} \leq \tau_{\mathbb{M}}^{-1} \log \left(4 C_{\mathbb{M}} \frac{L + \kappa}{\kappa} \frac{2 - \sqrt{q}}{q} \right) \quad \text{when} \quad \mu > 0,$$

$$T_{k+1} \leq \tau_{\mathbb{M}}^{-1} \log \left(4 C_{\mathbb{M}} \frac{L + \kappa}{\kappa} (k + 2)^2 \right) \quad \text{when} \quad \mu = 0.$$

Remark 3.3

This is Corollary 16 in Lin et al. [10] but phrased in our notations.

Total complexity under C2

The total number of iteration $N_{\mathbb{M}}$ experienced by $\mathbb{M}$ for $\mathbb{A}$ to produce x_k such that $F(x_k) - F^* \leq \epsilon$ when using termination criterion C2, warm start and error sequence strategies in Theorem 16 has:

1. $N_{\mathbb{M}} \leq \widetilde{\mathcal{O}}\left(\tau_{\mathbb{M}}^{-1}\sqrt{\kappa/\epsilon}\log(\epsilon^{-1})\right)$ if $\mu = 0$,

2. $N_{\mathbb{N}} \leq \widetilde{\mathcal{O}}\left((\tau_{\mathbb{M}}\sqrt{q})^{-1}\log(\epsilon^{-1})\right)$ if $\mu > 0$.

In addition, termination criterion C2 has advantages:

1. Relative error sequence δ_k doesn't requite F^*.

2. Doesn't require bounded level set assumption for inner loop complexity bound.

3. Generalize to non-convex settings which was also the focus of Paquette's paper on Catalyst [14].

Comments on Catalyst from Necoara et al.

Necoara et al. [11] considered extending the linear convergence results for problem $f^* = \min_{x \in X} f(x)$ with $f : X \subseteq \mathbb{R}^n \to \mathbb{R}$ and X closed and convex. They characterize conditions:

1. Quasi-strong convexity,
2. Quadratic under approximation,
3. Quadratic gradient growth,
4. quadratic functional growth,
5. and Global error bound conditions.

Discussion about Catalyst future works is summarized here:

Future works

Extend linear convergence of accelerated gradient for variance reduced incremental methods such as the Catalyst Framework into a weaker setting.

Our question on Catalyst Acceleration

A key detail about Catalyst

The smoothness assumption is baked into Assumption 3.2, where it states that the $\mathbb{M}$ needs to have linear convergence rate.

All VRMs method in the literature requires Lipschitz smoothness and strong convexity to achieve linear convergence rate. No linear convergent first order algorithm exists for all strong convex function by Theorem 2.1.5 in Nesterov's book [13]. Hence, here is our question:

Our question

Solving inexact proximal point doesn't necessarily need smoothness assumption on F so what if we Assumption 3.2 is false?

Next slide we explain why the answer is not trivial and why it's interesting.

The answer to our question is nontrivial

Without smoothness assumption on F, the following considerations apply.

1. $\mathbb{A}$ optimizes inexact Moreau Envelope $\mathcal{M}^*_{\kappa^{-1}, F}(y)$ which is smooth and strongly convex if F is convex hence $\mathcal{O}(1/\sqrt{k})$ from Theorem 3.2.1 [13] doesn't apply for the total complexity.

2. Theorem 2.1.7 [13] doesn't apply because the lower complexity bound $\mathbb{A}$ is $\mathcal{O}(1/k^2)$ assume exact gradient evaluations.

3. Theorem 2.1.17 [13] (Linear rate lower bound for strong convexity and smoothness) doesn't apply on $\mathbb{M}$. Instead Theorem 3.2.5 applies and the lower bound is $\mathcal{O}(1/k)$.

Subgradient descent with Polyak step size reaches the lower bound of worst case of Strongly convex, strongly convex and smooth. See Hazan et al. [7].

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.

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

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

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

References II

E. Hazan and S. Kakade, *Revisiting the Polyak step size*, Aug. 2022.

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.

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, 2015.

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

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.

Y. Nesterov, *Gradient methods for minimizing composite functions*, Mathematical Programming, 140 (2013), pp. 125–161.

References III

Y. NESTEROV, *Lectures on Convex Optimization*, vol. 137 of Springer Optimization and Its Applications, Springer International Publishing, 2018.

C. PAQUETTE, H. LIN, D. DRUSVYATSKIY, J. MAIRAL, AND Z. HARCHAOUI, *Catalyst for gradient-based nonconvex optimization*, in Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, PMLR, Mar. 2018, pp. 613–622.

E. K. RYU AND W. YIN, *Large-scale Convex Optimization: Algorithms & Analyses via Monotone Operators*, Cambridge University Press, 2022.

M. SCHMIDT, N. LE ROUX, AND F. BACH, *Minimizing finite sums with the stochastic average gradient*, Mathematical Programming, 162 (2017), pp. 83–112.

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

L. XIAO AND T. ZHANG, *A proximal stochastic gradient method with progressive variance reduction*, SIAM Journal on Optimization, 24 (2014), pp. 2057–2075.