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

Hongda Li

University of British Columbia Okanagan

January 26, 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

Notations and preliminaries

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

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

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 $k \ge 0$, 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 $(x_k)_{k\ge 0}$ such that for all $k \ge 0$ it has $F(x_k) \le \phi_k^* := \min_x \phi_k(x)$.
2. There exists a sequence of $(\alpha_k)_{k\ge 0}$ where $\alpha_k \in (0,1)\ \forall k \ge 0$ such that for all $x \in \mathbb{R}^n$ it has $\phi_{k+1}(x) - \phi_k(x) \le -\alpha_k(\phi_k(x) - F(x))$.

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:
>
> $$\text{find } x_k \approx \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ F(x) + (\kappa/2)\|x - y_{k-1}\|^2 \right\},$$
>
> $$\text{find } \alpha_k \in (0, 1) \text{ such that } \alpha_k^2 = (1 - \alpha_k)\alpha_{k-1}^2 + (\mu/(\mu + \kappa))\alpha_k,$$
>
> $$y_k = x_k + \frac{\alpha_{k-1}(1 - \alpha_{k-1})}{\alpha_{k-1}^2 + \alpha_k}(x_k - x_{k-1}).$$

We will return to this in the later slides.

Introducing Catalyst Part II

Catalyst by Lin, et al. [9, 8] has the theoretical and practical importance:

1. It's an early attempt at putting 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 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 $(\alpha_k)_{k\ge 0}, (\rho_k)_{k \ge 0}$ given by Definition 5. Initialize any x_1, v_1 in $\mathbb{R}^n$. For $k \ge 1$, the algorithm generates sequence of vector iterates $(y_k, v_{k+1}, x_{k+1})_{k\ge 1}$ 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 [11], 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. [7] 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 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 $L/2\|x^+ - y\|^2 < D_f(x^+, y)$ **do**
    $L := 2L;$
    $x^+ = [I + L^{-1}\partial g](y_k - L^{-1}\nabla f(y_k));$
  end while
  $x_{k+1} := x^+;$
  $\alpha_{k+1} := (1/2)\left(\mu/L - \alpha_k^2 + \sqrt{(\mu/L - \alpha_k^2)^2 + 4\alpha_k^2}\right);$
  $\theta_{k+1} := \alpha_k(1 - \alpha_k)/(\alpha_k^2 + \alpha_{k+1});$
  $y_{k+1} := x_{k+1} + \theta_{k+1}(x_{k+1} - x_k);$
  Compute: $f(y_{k+1})$
  $\mu := (1/2)(2D_f(y_{k+1}, y_k)/\|y_{k+1} - y_k\|^2) + (1/2)\mu;$
end for

Simple quadratic optimizations

Our metric is called normalized optimality gap:

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

This is our first numerical experiment is:

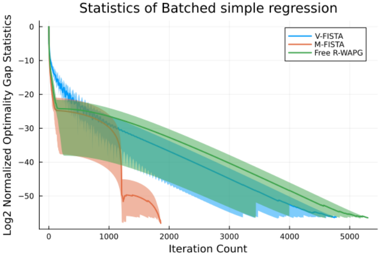
$$F(x) = (1/2)\langle x, Ax \rangle.$$

Our setup has:

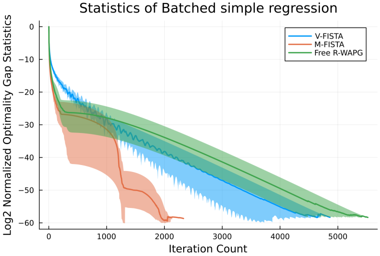
1. Same initial guess shared by FR-WAPG from us, and M-FISTA, V-FISTA in (10.7.7, 10.7.6) by Beck [2]. It's repeated for 30 different random initial guesses.
2. Min, max, and median of δ_k measured for all iterative method.
3. μ, L were given in prior to produce diagonal matrix $A = \text{diag}(0, \mu + (L-\mu)(N-1)^{-1}, \mu + 2(L-\mu)(N-1)^{-1}, \cdots, \mu + (N-2)(L-\mu)^{-1}, L)$, but 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.

μ estimation graph by FR-WAPG

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

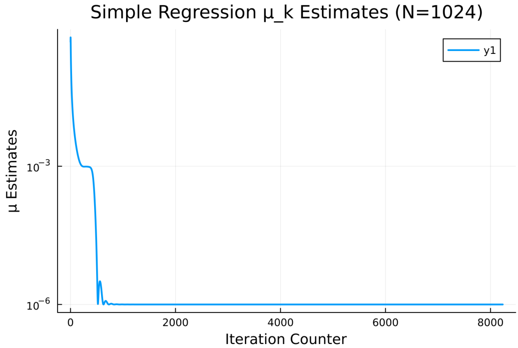


Figure: $N = 1024$, the μ estimates produced by Algorithm 1 (R-WAPG) is recorded.

LASSO numerical experiment

Tibshirani [13] proposed:

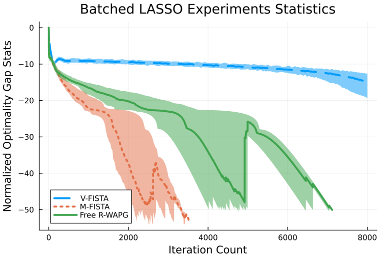
$$\min_x \left\{ \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1 \right\}.$$

Our setup:

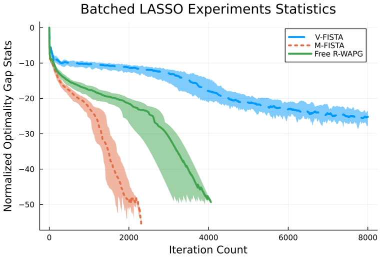
1. M, N are constants. $A \in \mathbb{R}^{M\times N}$ ahs i.i.d random entries from a standard normal distribution.
2. L, μ, are estimated by A by $\mu = 1/\|(A^T A)^{-1}\|$ and $L = \|A^T A\|$.
3. The synthetic solution is $x^+ = [1\ -1\ 1\ \cdots]^T \in \mathbb{R}^N$ and $b = Ax^+ \in \mathbb{R}^M$.
4. $x_0 \in \mathbb{R}^N$ is the initial guess with i.i.d randomed variable from a standard normal distributions. 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^*.

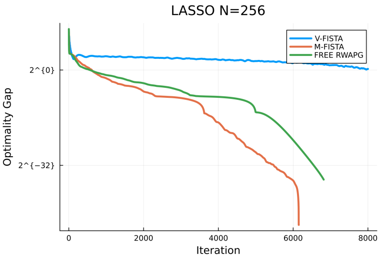


(b) LASSO experiment with $M = 64, N = 128$. Plots of minimum, maximum, and median δ_k with estimated F^*.

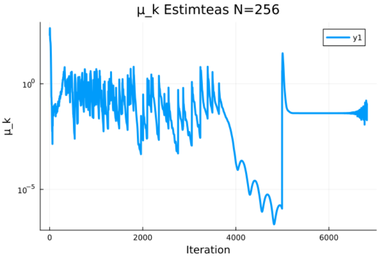
Figure: LASSO experiments.

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 $\mu = 7.43236362761395 8 \times 10^{-18}$ and $L = 2321.737206983643$.

Nesterov's idea of strong convexity transfer

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

1. In Nesterov's 2013 paper [10], 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.

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:

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

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:

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

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

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 $\mathbb{M}$. Major examples of VRMs include SVRG by Xiao, Zhang [14], Finito by Defazio et al. [6], SAG by Schmidt et al. [12], 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 $A > 0$, and it has:

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

Outer loop complexity, strong convex case

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

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

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

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

Outer loop complexity, convex but not strongly convex case

Theorem 14

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

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

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

Proposition 3.1 (Inner loop complexity strongly convex)

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 in μ, κ and other constants.

Proposition 3.2 (Inner loop, context but not strongly convex)

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 $z_{0,k} = x_{k-1}$,
3. 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.

Aggregated complexity

We count m, the number of iteration experienced by $\mathbb{M}$ for the k th 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:

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

2. If $\mu = 0$, using Proposition 3.2, Theorem 14 it has

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

So:

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

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 $\alpha_0 \in (0, 1]$. 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 $x_k \in \widetilde{\mathcal{J}}^{\delta_k}_{\kappa^{-1}F} y_{k-1}$,

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

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

Outerloop 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 [9]. For a precise description of the upper bound, see Theorem 8.

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

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.

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

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

References III

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.