Non-Nesterov Acceleration Methods for Making Gradients Small in Convex Minimization and Convex-Concave Minimax Optimization

Ernest K. Ryu

Department of Mathematical Sciences
Seoul National University

One World Optimization Seminar
October 11, 2021

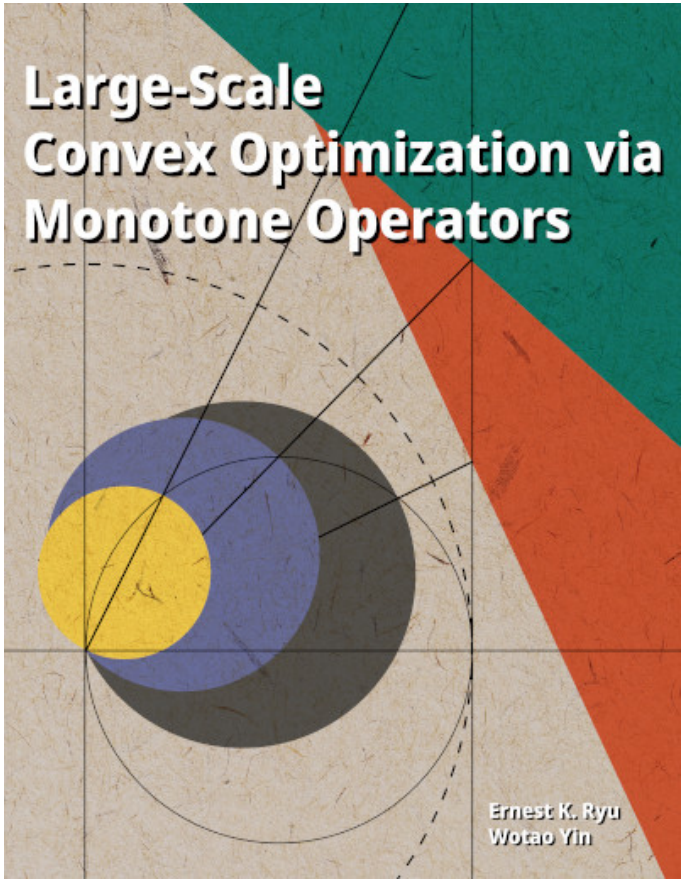
Update on me

Last year, I moved to Seoul National Univesity, Korea.

Prior to moving, my primary research area was monotone operator theory. (Also, Wotao and I have finally finished our book on monotone operator theory and splitting methods.)

Since, I've started to work on machine learning and acceleration.

Today, I'll share my recent work on acceleration.



Acceleration of first-order convex minimization

Consider

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x)$$

where f is L-smooth convex. Gradient descent

$$x_{k+1} = x_k - \frac{1}{L}f(x_k)$$

converge with the rate $f(x_k) - f_\star \leq \mathcal{O}(1/k)$.
Nesterov's celebrated accelerated gradient method (AGM)

$$y_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$
$$x_{k+1} = y_{k+1} + \frac{k-1}{k+2}(y_{k+1} - y_k)$$

converges with the accelerated rate $f(x_k) - f_\star \leq \mathcal{O}(1/k^2)$.

Question) Can we accelerate methods for other setups?

Outline

Acceleration for smooth convex-concave minimax optimization

Acceleration for monotone inclusions and fixed-point iterations

Acceleration for making gradients small in smooth convex minimization

Smooth convex-concave minimax optimization

We consider

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\ \underset{\mathbf{y}\in\mathbb{R}^m}{\text{maximize}}\quad \mathbf{L}(\mathbf{x},\mathbf{y}),$$

where $\mathbf{L}$ is convex-concave and R-smooth. Recently, minimax optimization has gained popularity in machine learning.

$(\mathbf{x}^\star,\mathbf{y}^\star)$ solves the minimax problem if it is a saddle point, i.e., if

$$\mathbf{L}(\mathbf{x}^\star,\mathbf{y}) \le \mathbf{L}(\mathbf{x}^\star,\mathbf{y}^\star) \le \mathbf{L}(\mathbf{x},\mathbf{y}^\star), \qquad \forall \mathbf{x}\in\mathbb{R}^n,\ \mathbf{y}\in\mathbb{R}^m.$$

Saddle operator is

$$\mathbf{G}(\mathbf{x},\mathbf{y}) \triangleq \begin{bmatrix} \nabla_{\mathbf{x}}\mathbf{L}(\mathbf{x},\mathbf{y}) \\ -\nabla_{\mathbf{y}}\mathbf{L}(\mathbf{x},\mathbf{y}) \end{bmatrix}.$$

$\mathbf{L}$ is R-smooth of $\mathbf{G}$ is R-Lipschitz continuous. $\mathbf{z}=(\mathbf{x},\mathbf{y})$ is a saddle point of $\mathbf{L}$ if and only if $\mathbf{G}(\mathbf{z})=0$.

Classical results in minimax optimization

Analogue of gradient descent (simultaneous gradient descent-ascent)

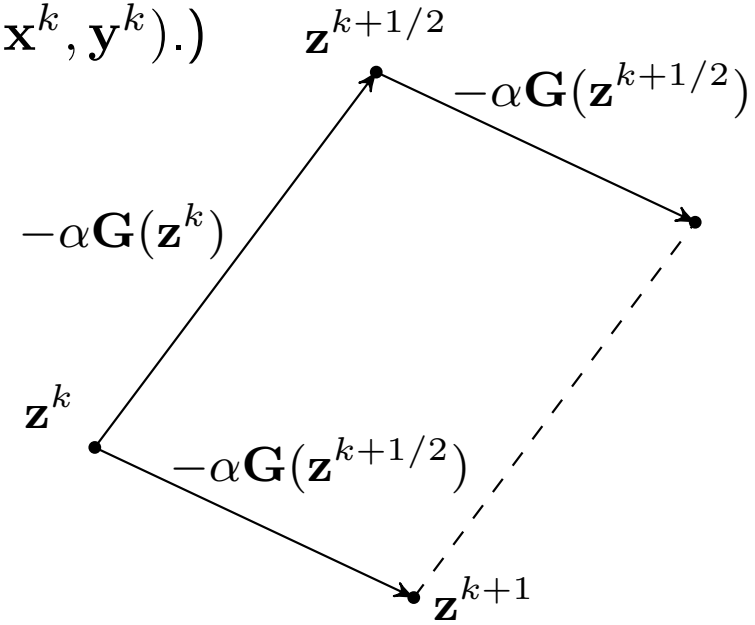
$$\mathbf{z}^{k+1} = \mathbf{z}^k - \alpha\, \mathbf{G}(\mathbf{z}^k),$$

does not converge in general. (Write $\mathbf{z}^k = (\mathbf{x}^k, \mathbf{y}^k)$.)

Extragradient (EG) algorithm[1]

$$\mathbf{z}^{k+1/2} = \mathbf{z}^k - \alpha\, \mathbf{G}(\mathbf{z}^k)$$
$$\mathbf{z}^{k+1} = \mathbf{z}^k - \alpha\, \mathbf{G}(\mathbf{z}^{k+1/2})$$

does converge.



[1] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. 1976.

EG is optimal

Duality gap naturally generalizes function value in convex minimization.

Theorem (Informal[2])

The averaged iterates of EG satisfy

$$\underbrace{\max_{\mathbf{y}\in Y} \mathbf{L}(\overline{\mathbf{x}}^k, \mathbf{y}) - \min_{\mathbf{x}\in X} \mathbf{L}(\mathbf{x}, \overline{\mathbf{y}}^k)}_{:=\textit{duality gap}} \leq \mathcal{O}\left(\frac{R\|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{k}\right).$$

Theorem (Informal[3])

Complexity lower bound for first-order gradient methods:

$$\textit{duality gap } (\mathbf{x}^k, \mathbf{y}^k) \geq \Omega\left(\frac{R\|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{k}\right).$$

---

[2] A. Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. 2004.

So is acceleration possible?

EG already optimal for duality gap. Further acceleration is impossible.

So is acceleration possible?

EG already optimal for duality gap. Further acceleration is impossible.

Or is it?

What if we change the optimality measure?

Duality gap, as an optimality measure, has drawbacks:

- Does not generalize to the non-convex-non-concave setup.
- Cannot be measured throughout the algorithm.

Optimality measures and gradient norm

What if we consider the squared gradient norm

$$\|\mathbf{G}(\mathbf{z})\|^2 = \|\nabla \mathbf{L}(\mathbf{z})\|^2$$

as the optimality measure?

Theorem (Informal)

EG and several other known methods exhibit the rate

$$\min_{i=0,\ldots,k} \|\nabla \mathbf{L}(\mathbf{z}^i)\|^2 \leq \mathcal{O}\left(\frac{R^2\|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{k}\right).$$

We can do better.

Main Results: Optimal acceleration for gradient norm

Contribution 1. Present *Extra Anchored Gradient (EAG)* with rate

$$\|\nabla \mathbf{L}(z^k)\|^2 \leq \mathcal{O}\left(\frac{R^2\|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{k^2}\right).$$

Contribution 2. Establish EAG's optimality with matching lower bound

$$\|\nabla \mathbf{L}(\mathbf{z}^k)\|^2 \geq \Omega\left(\frac{R^2\|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{k^2}\right).$$

Extra anchored gradient (EAG) algorithm

General form of EAG:

$$\mathbf{z}^{k+1/2} = \mathbf{z}^k + \frac{1}{k+2}(\mathbf{z}^0 - \mathbf{z}^k) - \alpha_k \mathbf{G}(\mathbf{z}^k)$$
$$\mathbf{z}^{k+1} = \mathbf{z}^k + \frac{1}{k+2}(\mathbf{z}^0 - \mathbf{z}^k) - \alpha_k \mathbf{G}(\mathbf{z}^{k+1/2})$$

$\alpha_k > 0$ are *step-sizes* and $\frac{1}{k+2}$ are *anchoring coefficients*.

Anchor term pulls $\mathbf{z}^k$ towards the initial point $\mathbf{z}^0$.

EAG-C

EAG with constant step-size (EAG-C):

$$\mathbf{z}^{k+1/2} = \mathbf{z}^k + \frac{1}{k+2}(\mathbf{z}^0 - \mathbf{z}^k) - \alpha \mathbf{G}(\mathbf{z}^k)$$
$$\mathbf{z}^{k+1} = \mathbf{z}^k + \frac{1}{k+2}(\mathbf{z}^0 - \mathbf{z}^k) - \alpha \mathbf{G}(\mathbf{z}^{k+1/2}),$$

where $\alpha > 0$ is fixed.

Theorem

With $\alpha = \frac{1}{8R}$, EAG-C exhibits the rate

$$\|\nabla \mathbf{L}(\mathbf{z}^k)\|^2 \leq \frac{260R^2\|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{(k+1)^2}.$$

EAG-C is simple, but analysis is very complicated. Constant is large as stepsize α is restrictive.

EAG-V

EAG with varying step-size (EAG-V):

$$\mathbf{z}^{k+1/2} = \mathbf{z}^k + \frac{1}{k+2}(\mathbf{z}^0 - \mathbf{z}^k) - \alpha_k \mathbf{G}(\mathbf{z}^k)$$
$$\mathbf{z}^{k+1} = \mathbf{z}^k + \frac{1}{k+2}(\mathbf{z}^0 - \mathbf{z}^k) - \alpha_k \mathbf{G}(\mathbf{z}^{k+1/2}),$$

where $\alpha_0 \in \left(0, \frac{1}{R}\right)$ and

$$\alpha_{k+1} = \frac{\alpha_k}{1 - \alpha_k^2 R^2}\left(1 - \frac{(k+2)^2}{(k+1)(k+3)}\alpha_k^2 R^2\right).$$

Theorem

With $\alpha_0 = \frac{0.618}{R}$, EAG-V exhibits the rate

$$\|\nabla \mathbf{L}(\mathbf{z}^k)\|^2 \leq \frac{27R^2\|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{(k+1)(k+2)}.$$

EAG-V is complicated, but analysis is simple. Constant is better as larger stepsizes α_k accomodated.

Proof outline

Theorem (Lyapunov analysis)

There exists a sequence $A_k = \Theta(k^2)$ *such that*

$$V_k \triangleq A_k \|\mathbf{G}(\mathbf{z}^k)\|^2 + (k+1)\langle \mathbf{G}(\mathbf{z}^k), \mathbf{z}^k - \mathbf{z}^0 \rangle$$

is nonincreasing in $k \geq 0$.

Proof outline

Using the Lyapunov function, we have

$$\begin{aligned} A_0 R^2 \|\mathbf{z}^0 - \mathbf{z}^\star\|^2 &\geq A_0 \|\mathbf{G}(\mathbf{z}^0)\|^2 = V_0 \geq \cdots \geq V_k \\ &= A_k \|\mathbf{G}(\mathbf{z}^k)\|^2 + (k+1)\langle \mathbf{G}(\mathbf{z}^k), \mathbf{z}^k - \mathbf{z}^0 \rangle \\ &\geq A_k \|\mathbf{G}(\mathbf{z}^k)\|^2 + (k+1)\langle \mathbf{G}(\mathbf{z}^k), \mathbf{z}^\star - \mathbf{z}^0 \rangle. \end{aligned}$$

With Young's inequality, we get

$$\left(A_0 R^2 + \frac{(k+1)^2}{2A_k} \right) \|\mathbf{z}^0 - \mathbf{z}^\star\|^2 \geq \frac{A_k}{2} \|\mathbf{G}(\mathbf{z}^k)\|^2.$$

Since $A_k = \Theta(k^2)$, we conclude $\|\mathbf{G}(\mathbf{z}^k)\|^2 \leq \mathcal{O}\left(\frac{R^2 \|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{k^2} \right)$. □

EAG is optimal up to a constant

Theorem

Let $k \geq 0$ and $n \geq k+2$. Then there exists an R-smooth saddle function $\mathbf{L}$ on $\mathbb{R}^n \times \mathbb{R}^n$ satisfying

$$\|\nabla \mathbf{L}(\mathbf{z}^k)\|^2 \geq \frac{R^2 \|\mathbf{z}^0 - \mathbf{z}^\star\|^2}{(k+1)^2}$$

for any $\mathbf{z}^0 \in \mathbb{R}^n \times \mathbb{R}^n$ and any iterative algorithm satisfying

$$\mathbf{x}^i \in \mathbf{x}^0 + \operatorname{span}\{\nabla_{\mathbf{x}} \mathbf{L}(\mathbf{x}^0, \mathbf{y}^0), \ldots, \nabla_{\mathbf{x}} \mathbf{L}(\mathbf{x}^{i-1}, \mathbf{y}^{i-1})\}$$
$$\mathbf{y}^i \in \mathbf{y}^0 + \operatorname{span}\{\nabla_{\mathbf{y}} \mathbf{L}(\mathbf{x}^0, \mathbf{y}^0), \ldots, \nabla_{\mathbf{y}} \mathbf{L}(\mathbf{x}^{i-1}, \mathbf{y}^{i-1})\}$$

for $i = 1, \ldots, k$.

The algorithm class contains both simultaneous and alternating gradient descent-ascent methods.

Worst-case saddle function construction

Worst-case bilinear saddle function:

$$\mathbf{L}(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\intercal \mathbf{A}\mathbf{y} - \mathbf{b}^\intercal \mathbf{x} - \mathbf{b}^\intercal \mathbf{y}$$

where $\mathbf{A} \in \mathbb{S}^{n \times n}$ and $\mathbf{b} \in \mathbb{R}^n$.

Note that $[\nabla \mathbf{L}(\mathbf{x}, \mathbf{y}) = 0] \Leftrightarrow [\mathbf{A}\mathbf{x} = \mathbf{b} \text{ and } \mathbf{A}\mathbf{y} = \mathbf{b}]$.

Since $\mathbf{A}$ is symmetric, when $\mathbf{x}^0 = \mathbf{y}^0 = 0$, the span conditions

$$\mathbf{x}^k \in \mathbf{x}^0 + \operatorname{span}\{\nabla_{\mathbf{x}} \mathbf{L}(\mathbf{x}^0, \mathbf{y}^0), \ldots, \nabla_{\mathbf{x}} \mathbf{L}(\mathbf{x}^{k-1}, \mathbf{y}^{k-1})\}$$
$$\mathbf{y}^k \in \mathbf{y}^0 + \operatorname{span}\{\nabla_{\mathbf{y}} \mathbf{L}(\mathbf{x}^0, \mathbf{y}^0), \ldots, \nabla_{\mathbf{y}} \mathbf{L}(\mathbf{x}^{k-1}, \mathbf{y}^{k-1})\}.$$

reduce to

$$\mathbf{x}^k, \mathbf{y}^k \in \mathcal{K}_{k-1}(\mathbf{A}; \mathbf{b}) \stackrel{\Delta}{=} \operatorname{span}\{\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^{k-1}\mathbf{b}\}$$

($\mathcal{K}_{k-1}(\mathbf{A}; \mathbf{b})$ is the $(k-1)$th Krylov subspace.)

Nemirovsky's lower bound for $\mathbf{Ax} = \mathbf{b}$

Key idea: bilinear minimax problems generalize $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{A} \in \mathbb{S}^{n \times n}$, so the following lower bound applies to our setup.

Lemma (Nemirovsky[3])

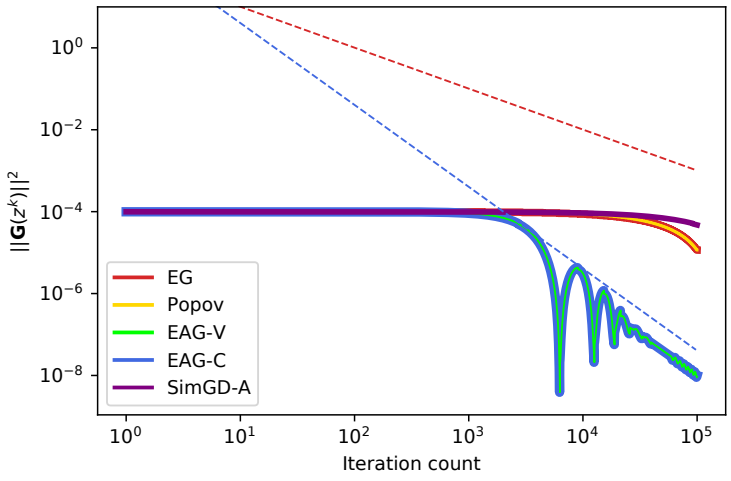
Let $R > 0$, $k \geq 0$ and $n \geq k + 2$. Then there exists a $\mathbf{A} \in \mathbb{S}^{n \times n}$ such that $\|\mathbf{A}\| \leq R$ and $0 \neq \mathbf{b} \in \mathcal{R}(\mathbf{A})$ satisfing

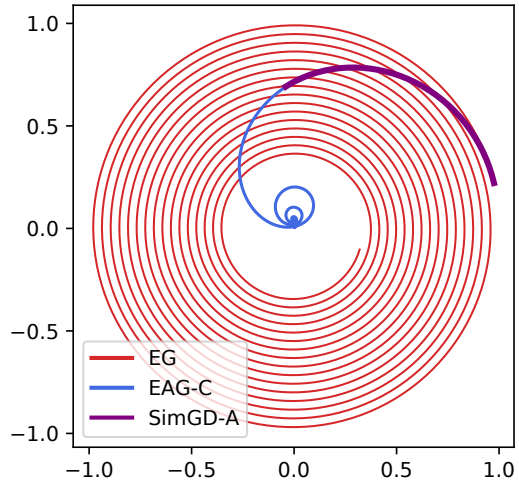
$$\|\mathbf{Ax} - \mathbf{b}\|^2 \geq \frac{R^2 \|\mathbf{x}^\star\|^2}{(k+1)^2}$$

for all $\mathbf{x} \in \mathcal{K}_{k-1}(\mathbf{A}; \mathbf{b})$, where $\mathbf{x}^\star$ is the minimum norm solution to $\mathbf{Ax} = \mathbf{b}$.

[3] Nemirovsky. Information-based complexity of linear operator equations. *J. Complexity*, 1992.

Experiments





(Left) Plots of $\|\nabla \mathbf{L}(\mathbf{z}^k)\|^2$ versus iteration count. Dashed lines indicate theoretical upper bounds. We observe the $\mathcal{O}(1/k^2)$ rate.

(Right) Comparison of trajectories. The anchoring mechanism dampens cycling behavior.

Summary

With EAG and a matching lower bound, we establish the optimal accelerated $\mathcal{O}(1/k^2)$ complexity on the squared gradient magnitude for smooth convex-concave minimax problems.

Reference:

T. Yoon and E. K. Ryu, Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm, *ICML*, 2021.

Outline

Acceleration for smooth convex-concave minimax optimization

Acceleration for monotone inclusions and fixed-point iterations

Acceleration for making gradients small in smooth convex minimization

Monotone inclusion problem

Consider the problem

$$\underset{x\in\mathbb{R}^n}{\text{find}} \quad 0 \in \mathbb{A}x,$$

where $\mathbb{A}$ is maximal monotone.

Proximal point method:

$$x^{k+1} = \mathbb{J}_{\mathbb{A}} x^k$$

exhibits the rate

$$\|x^k - \mathbb{J}_{\mathbb{A}} x^k\|^2 \leq \mathcal{O}\left(\frac{1}{k}\right).$$

Accelerated proximal point method

Accelerated proximal point method (APPM):

$$y^{k+1} = \mathbb{J}_{\mathbb{A}} x^k$$
$$x^{k+1} = y^{k+1} + \frac{k}{k+2}(y^{k+1} - y^k) - \frac{k}{k+2}(y^k - x^{k-1}),$$

where $y^0 = x^0$.

Exhibits the rate

$$\|x^k - \mathbb{J}_{\mathbb{A}} x^k\|^2 \leq \mathcal{O}\left(\frac{1}{k^2}\right).$$

Kim, Accelerated proximal point method for maximally monotone operators, *MPA*, 2021.

Fixed-point problem

Consider the problem

$$\operatorname*{find}_{x \in \mathbb{R}^n} \quad x = \mathbb{T}x,$$

where $\mathbb{T}\colon \mathbb{R}^n \to \mathbb{R}^n$ is nonexpansive.

Krasnosel'skiĭ–Mann iteration:

$$x^{k+1} = \frac{1}{2}x^k + \frac{1}{2}\mathbb{T}x^k$$

exhibits the rate

$$\|x^k - \mathbb{T}x^k\|^2 \leq \mathcal{O}\left(\frac{1}{k}\right).$$

Optimized Halpern method

Optimized Halpern method (OHM):

$$x^{k+1} = \frac{1}{k+2}x^0 + \frac{k+1}{k+2}\mathbb{T}x^k.$$

Exhibits the rate

$$\|x^k - \mathbb{T}x^k\|^2 \leq \mathcal{O}\left(\frac{1}{k^2}\right).$$

Lieder, On the convergence rate of the Halpern-iteration, *OPTL*, 2021.

APPM $\cong$ OHM

The two independent discoveries, APPM and OHM, are equivalent.

Kim and Lieder discovered these methods with a computer-assisted methodology, the performance estimation problem[†]. The presented proofs are verifiable but arguably difficult to understand.

[†] Drori and Teboulle, Performance of first-order methods for smooth convex minimization: a novel approach. *MPA*, 2014.

[†] Taylor, Hendrickx, and Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods, *MPA*, 2017.

Accelerated rates of APPM and OHM

Theorem

APPM/OHM exhibits the rate

$$\|x^{k-1} - \mathbb{J}_{\mathbb{A}} x^{k-1}\|^2 \leq \frac{\|x^0 - x^\star\|^2}{k^2}$$

for $k = 1, 2, \ldots$.

Equivalently,

$$\|\mathbb{T} x^{k-1} - x^{k-1}\|^2 \leq \frac{4\|x^0 - x^\star\|^2}{k^2}.$$

Accelerated rates of APPM and OHM

Proof. Define $\tilde{\mathbb{A}}y^k = x^{k-1} - y^k$, which implies $\tilde{\mathbb{A}}y^k \in \mathbb{A}y^k$. Define

$$V^k = \frac{k^2}{2}\|\tilde{\mathbb{A}}y^k\|^2 + k\langle \tilde{\mathbb{A}}y^k, y^k - x^\star\rangle + \frac{1}{2}\|k\tilde{\mathbb{A}}y^k - (x^0 - x^\star)\|^2$$

for $k = 0, 1, \ldots$. Then

$$V^{k+1} - V^k = -k(k+1)\langle \tilde{\mathbb{A}}y^{k+1} - \tilde{\mathbb{A}}y^k, y^{k+1} - y^k\rangle \leq 0.$$

Conclude

$$\frac{k^2}{2}\|\tilde{\mathbb{A}}y^k\|^2 \leq V^k \leq V^0 = \frac{1}{2}\|x^0 - x^\star\|^2.$$

□

This Lyapunov proof is due to: Park and Ryu, Exact Optimal Accelerated Complexity for Fixed-Point Iterations, *upcoming*, 2021.

A proof with a similar structure was also presented in: Diakonikolas, Halpern iteration for near-optimal and parameter-free monotone inclusion and strong solutions to variational inequalities, *COLT*, 2020.

Optimality of APPM and OHM

APPM/OHM are exactly optimal; they have an exact matching complexity lower bound.

Theorem (Informal)

Given k, there exists an operator $\mathbb{A}$ such that

$$\|x^{k-1} - \mathbb{J}_{\mathbb{A}} x^{k-1}\|^2 \geq \frac{\|x^0 - x^\star\|^2}{k^2}$$

for any algorithm satisfying the span condition.

Accelerating Picard and Banach

When $\mathbb{T}$ is contractive, i.e., γ-Lipschitz with $\gamma < 1$, the classical iteration

$$x^{k+1} = \mathbb{T}x^k$$

exhibits the rate $\mathcal{O}(\gamma^k)$.

Theorem (Informal)

Using a mechanism analogous to APPM/OHM, we can accelerate the classical fixed-point iteration in the contractive (strongly monotone) setup. This accelerated rate is exactly optimal.

Acceleration under quasi-uniform monotonicity

$\mathbb{A} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is *quasi-uniformly monotone with parameters* $\mu > 0$ *and* $\alpha > 1$ if it is monotone and

$$\langle \mathbb{A}x, x - x^\star \rangle \geq \mu \|x - x^\star\|^{\alpha+1}$$

for any $x \in \mathbb{R}^n$ and $x^\star \in \mathrm{Zer}\mathbb{A}$. ($\alpha = 1$ and $\alpha = \infty$ respectively correspond to strong and plain monotonicity.)

Theorem

Under quasi-uniformly monotonicity, PPM exhibits the rate

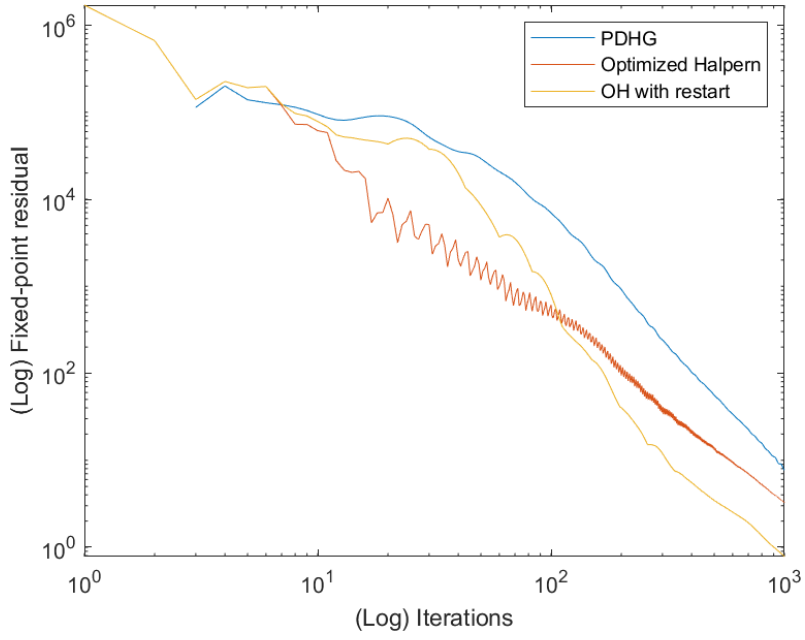
$$\|\mathbb{A}x_k\|^2 \leq \mathcal{O}\left(k^{-\frac{\alpha+1}{\alpha-1}}\right).$$

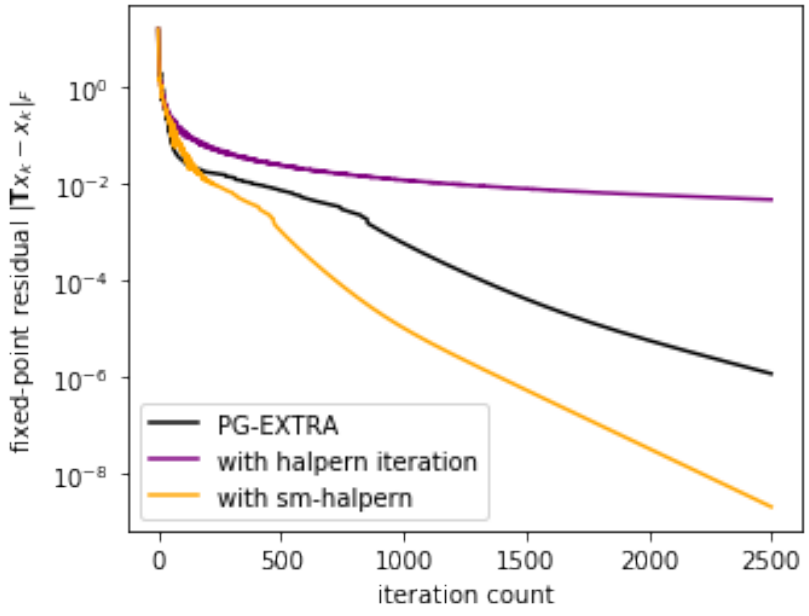
Theorem (Informal)

Under quasi-uniformly monotonicity, we can use a restarting scheme accelerate the rate to

$$\|\mathbb{A}x_k\|^2 \leq \mathcal{O}\left(k^{-\frac{2\alpha}{\alpha-1}}\right).$$

Experiments





(Left) Total variation CT reconstruction with PDHG. OHM with restart faster at later iterations.

(Right) Decentralized compressed sensing with PG-EXTRA. Faster linear convergence with method analogous to APPM/OHM.

Summary

The classical fixed-point iterations are suboptimal.

We present acceleration schemes for fixed-point iterations and provide matching complexity lower bounds.

Reference:
J. Park and E. K. Ryu, Exact optimal accelerated complexity for fixed-point iterations, *upcoming*, 2021.

Outline

Acceleration for smooth convex-concave minimax optimization

Acceleration for monotone inclusions and fixed-point iterations

Acceleration for making gradients small in smooth convex minimization

Is there a geometric structure of acceleration?

The many accelerated methods have been developed and analyzed with disparate techniques, without a unified framework. Is there a geometric structure of acceleration?

In this work, we identify the "parallel" and "collinear" structures of acceleration.

Using this insight, we better understand the acceleration of OGM-G and extended the acceleration to the prox-grad setup.

OGM-G: $\mathcal{O}((f(x_0) - f_\star)/K^2)$ rate

OGM-G:

$$x_{k+1} = x_k^+ + \frac{(\theta_k - 1)(2\theta_{k+1} - 1)}{\theta_k(2\theta_k - 1)}(x_k^+ - x_{k-1}^+) + \frac{2\theta_{k+1} - 1}{2\theta_k - 1}(x_k^+ - x_k)$$

where $x^+ = x - \frac{1}{L}\nabla f(x)$ and $\theta_k^2 - \theta_k = \theta_{k+1}^2$.

OGM-G exhibts the rate

$$\|\nabla f(x_K)\|^2 \leq \mathcal{O}((f(x_0) - f_\star)/K^2).$$

Discovered with a computer-assisted methodology. Original proof by KF was verifiable but arguably difficult to understand.

Kim and Fessler, Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions, *JOTA*, 2021.

FGM+OGM-G: $\mathcal{O}(\|x_0 - x_\star\|^2/K^4)$ rate

FGM+OGM-G: From x_0 run K iterations of FGM. Continue with OGM-G and run K iterations. Concatenated method exhibits the rate

$$\|\nabla f(x_{2K})\|^2 \leq \mathcal{O}(\|x_0 - x_\star\|^2/K^4)$$

FGM: $\mathcal{O}(1/K^2)$ rate on $\big(\|x_0 - x_\star\|^2 \mapsto f(x_K) - f(x_\star)\big)$.

OGM-G: $\mathcal{O}(1/K^2)$ rate on $\big(f(x_0) - f(x_\star) \mapsto \|\nabla f(x_K)\|^2\big)$.

FGM+OGM-G: $\mathcal{O}(1/K^4)$ rate on $\big(\|x_0 - x_\star\|^2 \mapsto \|\nabla f(x_{2K})\|^2\big)$.

Nesterov, Gasnikov, Guminov, and Dvurechensky, Primal–dual accelerated gradient methods with small-dimensional relaxation oracle, *Optimization Methods and Software*, 2020.

Nesterov's FGM

Nesterov's FGM:

$$x_{k+1} = x_k^+ + \frac{\theta_k - 1}{\theta_{k+1}}(x_k^+ - x_{k-1}^+),$$

where $y_0 = x_0$, $\theta_0 = 1$, and $\theta_{k+1}^2 - \theta_{k+1} = \theta_k^2$ for $k = 0, 1, \ldots$.

Equivalent form: with $z_0 = x_0$,

$$z_{k+1} = z_k - \frac{\theta_k}{L}\nabla f(x_k)$$
$$x_{k+1} = \left(1 - \frac{1}{\theta_{k+1}}\right) x_k^+ + \frac{1}{\theta_{k+1}} z_{k+1}.$$

Analysis of FGM

FGM's rate

$$f(x_{k-1}^+) - f_\star \leq \frac{2L\|x_0 - x_\star\|^2}{k^2} + o\left(\frac{1}{k^2}\right)$$

established through Lyapunov analysis: define

$$U_k = \theta_{k-1}^2 \left(f(x_{k-1}^+) - f_\star\right) + \frac{L}{2}\|z_k - x_\star\|^2$$

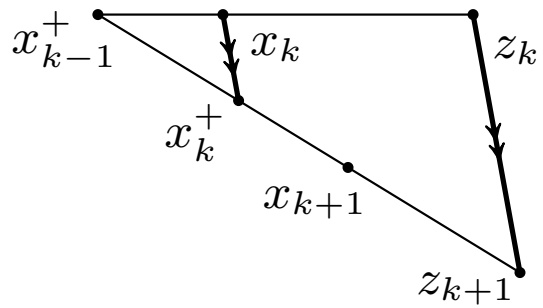
and show $U_k \leq \cdots \leq U_0$.

Nesterov, A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$, *Proceedings of the USSR Academy of Sciences*, 1983.

Parallel structure of FGM

Geometric observation. In FGM, $x_k^+ - x_k$ and $z_{k+1} - z_k$ are parallel.

Plane of iteration of FGM:



OGM

OGM:

$$x_{k+1} = x_k^+ + \frac{\theta_k - 1}{\theta_{k+1}}(x_k^+ - x_{k-1}^+) + \frac{\theta_k}{\theta_{k+1}}(x_k^+ - x_{k-1}^+)$$

for $k = 0, 1, \ldots,$ where $y_0 = x_0$.

Equivalent form: with $z_0 = x_0$,

$$z_{k+1} = z_k - \frac{2\theta_k}{L}\nabla f(x_k)$$
$$x_{k+1} = \left(1 - \frac{1}{\theta_{k+1}}\right) x_k^+ + \frac{1}{\theta_{k+1}} z_{k+1}.$$

Drori and Teboulle, Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 2014.

Kim and Fessler, Optimized first-order methods for smooth convex minimization, *Mathematical Programming*, 2016

Analysis of OGM

OGM's rate

$$f(x_{k-1}^+) - f_\star \leq \frac{L\|x_0 - x_\star\|^2}{k^2} + o\left(\frac{1}{k^2}\right)$$

established through Lyapunov analysis: define

$$U_k = 2\theta_k^2 \left(f(x_k) - f_\star - \frac{1}{2L}\|\nabla f(x_k)\|^2 \right) + \frac{L}{2}\|z_{k+1} - x_\star\|^2$$

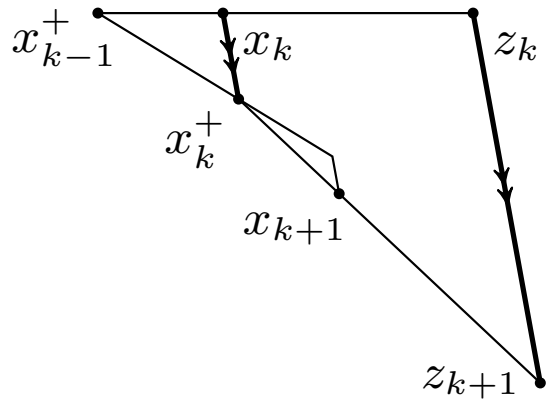
and show $U_k \leq \cdots \leq U_0$.

Park, Park, and Ryu, Factor-$\sqrt{2}$ acceleration of accelerated gradient methods, *arXiv*, 2021.

Parallel structure of OGM

Geometric observation. In OGM, $x_k^+ - x_k$ and $z_{k+1} - z_k$ are parallel.

Plane of iteration of OGM:



Parallel and collinear structure

Define this as the *parallel structure*.

For the strongly convex setup, define an analogous *collinear structure*.

Many of accelerated methods satisfy these structures: Nesterov's FGM, OGM, OGM-G, Nesterov's FGM in the strongly convex setup (SC-FGM), SC-OGM, TMM, non-stationary SC-FGM, ITEM, geometric descent, Güler's first and second accelerated proximal methods, and FISTA.

The parallel structure of OGM-G

For OGM-G, define z_k-sequence so that the parallel structure is safisfied.

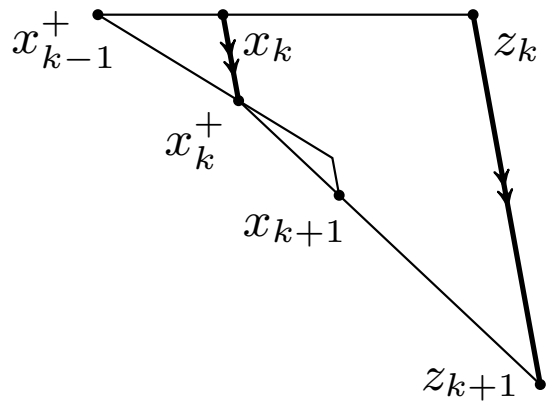
Equivalent form of OGM-G:

$$x_k = \frac{\theta_{k+1}^4}{\theta_k^4} x_{k-1}^+ + \left(1 - \frac{\theta_{k+1}^4}{\theta_k^4}\right) z_k$$

$$z_{k+1} = z_k - \frac{\theta_k}{L} \nabla f(x_k).$$

Now, $x_k^+ - x_k$ and $z_{k+1} - z_k$ are parallel.

Plane of iteration of OGM-G:



New analysis of OGM-G

We can now perform a Lyapunov analysis of OGM-G: define

$$U_k = \frac{1}{\theta_k^2}\left(\frac{1}{2L}\|\nabla f(x_K)\|^2 + \frac{1}{2L}\|\nabla f(x_k)\|^2 + f(x_k) - f(x_K) - \left\langle \nabla f(x_k), x_k - x_{k-1}^+ \right\rangle\right)$$
$$+ \frac{L}{\theta_k^4}\left\langle z_k - x_{k-1}^+, z_k - x_K^+ \right\rangle$$

and show $U_k \leq \cdots \leq U_0$.

A similar Lyapunov-type analysis was also presented in: Diakonikolas and Wang, Potential function-based framework for making the gradients small in convex and min-max optimization, *arXiv*, 2021.

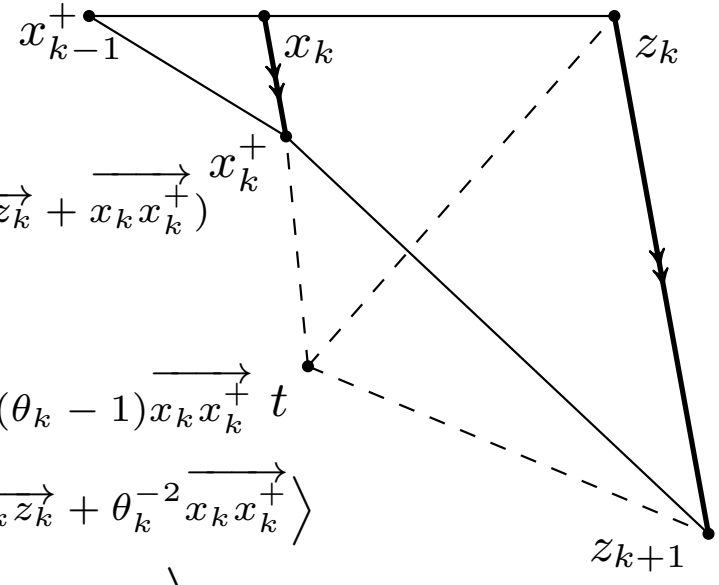
New analysis of OGM-G

The Lyapunov function U_k is obtained from cocoercivity inequalities and the parallel structure of OGM-G:

$$
\begin{aligned}
0 \geq\ & \frac{1}{\theta_{k+1}^2}\left(f(x_{k+1}) - f(x_k) - \langle \nabla f(x_{k+1}), x_{k+1} - x_k\rangle + \frac{1}{2L}\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2\right) \\
& + \left(\frac{1}{\theta_{k+1}^2} - \frac{1}{\theta_k^2}\right)\left(f(x_k) - f(x_K) - \langle \nabla f(x_k), x_k - x_K\rangle + \frac{1}{2L}\|\nabla f(x_k) - \nabla f(x_K)\|^2\right) \\
=\ & \frac{1}{\theta_{k+1}^2}\left(\frac{1}{2L}\|\nabla f(x_K)\|^2 + \frac{1}{2L}\|\nabla f(x_{k+1})\|^2 + f(x_{k+1}) - f(x_K) - \left\langle \nabla f(x_{k+1}), x_{k+1} - x_k^+\right\rangle\right) \\
& - \frac{1}{\theta_k^2}\left(\frac{1}{2L}\|\nabla f(x_K)\|^2 + \frac{1}{2L}\|\nabla f(x_k)\|^2 + f(x_k) - f(x_K) - \left\langle \nabla f(x_k), x_k - x_{k-1}^+\right\rangle\right) \\
& \underbrace{- \left\langle \nabla f(x_k), \theta_{k+1}^{-2} x_k^+ - \theta_k^{-2} x_{k-1}^+ - \left(\theta_{k+1}^{-2} - \theta_k^{-2}\right) x_K^+\right\rangle}_{:=T}.
\end{aligned}
$$

New analysis of OGM-G

The term T can be understood as follows. Define $\overrightarrow{uv} = v - u$. Then



$$
\begin{aligned}
\frac{1}{L}T &\overset{\text{(i)}}{=} \left\langle \overrightarrow{x_k x_k^+}, (\theta_{k+1}^{-2} - \theta_k^{-2}) t\overrightarrow{x_k^+} + \theta_k^{-2}\overrightarrow{x_{k-1}^+ x_k^+} \right\rangle \\
&\overset{\text{(ii)}}{=} \left\langle \overrightarrow{x_k x_k^+}, (\theta_{k+1}^{-2} - \theta_k^{-2})(\overrightarrow{tz_{k+1}} - \overrightarrow{z_k z_{k+1}} - \overrightarrow{x_k z_k} + \overrightarrow{x_k x_k^+}) \right. \\
&\qquad \left. + \theta_k^{-2}(\overrightarrow{x_{k-1}^+ x_k} + \overrightarrow{x_k x_k^+}) \right\rangle \\
&\overset{\text{(iii)}}{=} \left\langle \overrightarrow{x_k x_k^+}, (\theta_{k+1}^{-2} - \theta_k^{-2})\overrightarrow{tz_{k+1}} - (\theta_{k+1}^{-2} - \theta_k^{-2})(\theta_k - 1)\overrightarrow{x_k x_k^+} \right. \\
&\qquad \left. - \left(\theta_{k+1}^{-2} - \theta_k^{-2}\right)\overrightarrow{x_k z_k} + (2\theta_k - 1)\theta_{k+1}^{-4}\overrightarrow{x_k z_k} + \theta_k^{-2}\overrightarrow{x_k x_k^+} \right\rangle \\
&\overset{\text{(iv)}}{=} \left\langle \overrightarrow{x_k x_k^+}, \left(\theta_{k+1}^{-2} - \theta_k^{-2}\right)\overrightarrow{tz_{k+1}} + \theta_{k+1}^{-2}(\theta_k - 1)^{-1}\overrightarrow{x_k z_k} \right\rangle \\
&\overset{\text{(v)}}{=} \theta_{k+1}^{-4}\left\langle \overrightarrow{x_k^+ z_{k+1}} - \overrightarrow{x_k z_k}, \overrightarrow{tz_{k+1}} \right\rangle + \theta_{k+1}^{-4}\left\langle \overrightarrow{tz_{k+1}} - \overrightarrow{tz_k}, \overrightarrow{x_k z_k} \right\rangle \\
&\overset{\text{(vi)}}{=} \theta_{k+1}^{-4}\left\langle z_{k+1} - x_k^+, z_{k+1} - x_K^+ \right\rangle - \theta_k^{-4}\left\langle z_k - x_{k-1}^+, z_k - x_K^+ \right\rangle.
\end{aligned}
$$

Prox-grad setup

Consider the problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad F(x) := f(x) + g(x),$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and L-smooth g is proximable.

Prox-grad step notation:

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

FISTA-G

A novel method, FISTA-G:

$$x_{k+1} = x_k^{\oplus} + \frac{\varphi_{k+1} - \varphi_{k+2}}{\varphi_k - \varphi_{k+1}}(x_k^{\oplus} - x_{k-1}^{\oplus})$$

for $k = 0, 1, \ldots, K-1$, where $x_{-1}^{\oplus} := x_0$, $\varphi_{K+1} = 0$, $\varphi_K = 1$, and

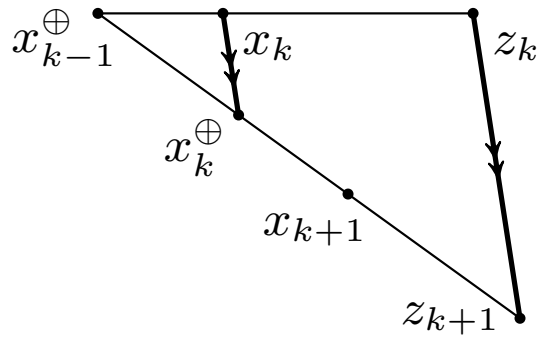
$$\varphi_k = \frac{\varphi_{k+2}^2 - \varphi_{k+1}\varphi_{k+2} + 2\varphi_{k+1}^2 + (\varphi_{k+1} - \varphi_{k+2})\sqrt{\varphi_{k+2}^2 + 3\varphi_{k+1}^2}}{\varphi_{k+1} + \varphi_{k+2}}$$

for $k = -1, 0, \ldots, K-1$.

Parallel structure of FISTA-G

Define $z_0 = x_0$, $z_k = \frac{\varphi_k}{\varphi_k - \varphi_{k+1}} x_k - \frac{\varphi_{k+1}}{\varphi_k - \varphi_{k+1}} x_{k-1}^{\oplus}$ for $k = 0, 1, \ldots, K$.
Then $x_k^+ - x_k$ and $z_{k+1} - z_k$ are parallel.

Plane of iteration of FISTA-G:



FISTA-G: $\mathcal{O}\left(\left(F(x_0) - F_\star\right)/K^2\right)$ rate

Theorem

FISTA-G's final iterate x_K exhibits the rate

$$\min \left\|\partial F(x_K^{\oplus})\right\|^2 \leq 4\left\|\tilde{\nabla}_L F(x_K)\right\|^2 \leq \frac{264L}{(K+2)^2}\left(F(x_0) - F_\star\right).$$

Proof outline. Define

$$U_k = \frac{2\varphi_{k-1}}{(\varphi_{k-1} - \varphi_k)^2}\left(\frac{1}{2L}\left\|\tilde{\nabla}_L F(x_k)\right\|^2 + F(x_k^{\oplus}) - F(x_K^{\oplus}) - \left\langle \tilde{\nabla}_L F(x_k), x_k - x_{k-1}^{\oplus}\right\rangle\right)$$
$$+ \frac{L}{\varphi_k}\left\langle z_k - x_{k-1}^{\oplus}, z_k - x_K^{\oplus}\right\rangle$$

and show $U_k \leq \cdots \leq U_0$. □

FISTA+FISTA-G: $\mathcal{O}((\|x_0 - x_\star\|^2)/K^4)$ rate

Corollary

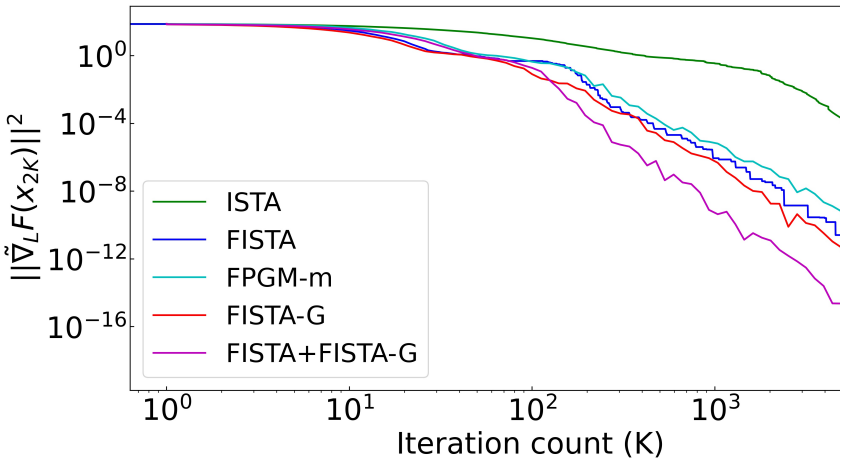
FISTA+FISTA-G's final iterate x_{2K} exhibits the rate

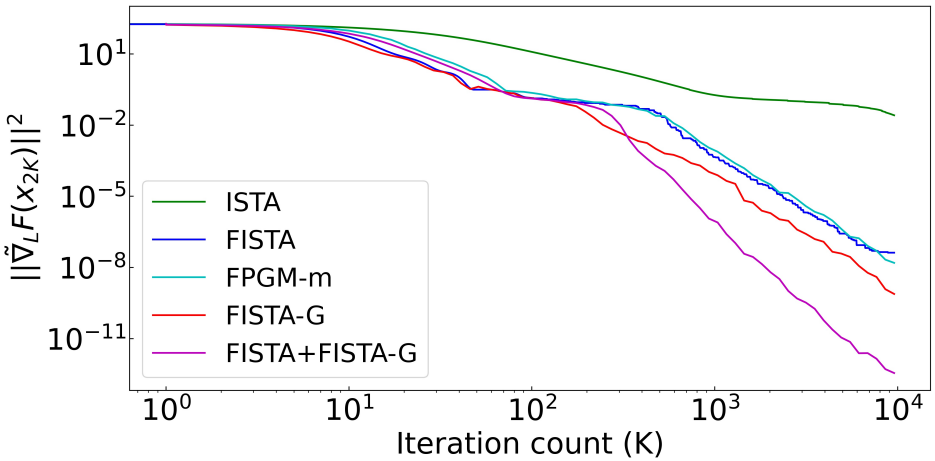
$$\min \|\partial F(x_{2K}^{\oplus})\|^2 \leq 4\|\tilde{\nabla}_L F(x_{2K})\|^2 \leq \frac{528L^2}{(K+2)^4}\|x_0 - x_\star\|^2.$$

Experiments

Compressed sensing experiments:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \|Ax-b\|^2+\lambda\|x\|_1, \qquad \underset{x\in\mathbb{R}^n}{\text{minimize}} \quad \|Ax-b\|^2+\lambda\|x\|_{\text{nuc}}.$$





Summary

We identify a geometric structure common among a wide range of accelerated first-order methods.

Using this geometric insight, we better understand the acceleration of OGM-G and extended the acceleration to the prox-grad setup.

Reference: J. Lee, C. Park, and E. K. Ryu, A geometric structure of acceleration and its role in making gradients small fast, *NeurIPS*, 2021.

Conclusion

The space of deterministic first-order convex optimization has a lot of exciting recent developments. With the aid the PEP[†], several new acceleration mechanisms have been discovered.

Open problem: Is there a common underlying structure to these acceleration mechanisms, despite their apparent differences?

[†] Drori and Teboulle, Performance of first-order methods for smooth convex minimization: a novel approach. *MPA*, 2014.

[†] Taylor, Hendrickx, and Glineur. Smooth strongly convex interpolation and exact worst-case performance of first-order methods, *MPA*, 2017.