On the Global R-linear Convergence of NAG Method and Beyond

Chenglong Bao

Yau Mathematical Sciences Center, Tsinghua University
Yanqi Lake Beijing Institute of Mathematical Sciences and Applications
State Key Laboratory of Membrane Biology, Tsinghua University

BU-Keio-Tsinghua workshop

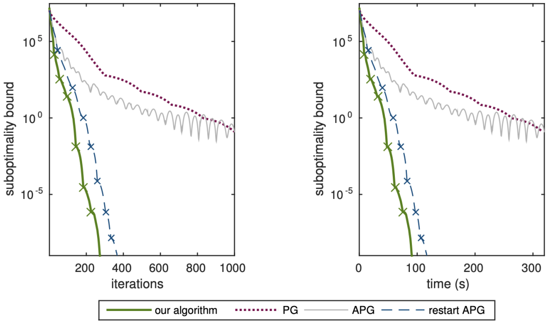
May 28, 2024

Outline

The starting point

Phase space tomography[1]: recover the coherence of a partially coherent light

- Algorithm: restarted accelerated gradient method
- Idea: mathematical explanations for the acceleration after restart
- Su et al. (2016) provides ODE perspective, but can not fully explain it



[1]SIIMS 2018; JOSA A 2017

Nonlinear convex optimization

Model

$$\min_x f(x) \tag{1}$$

The objective function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies

Assumptions on f

- L-**smooth**: $f \in C^1$ and

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n$$

- μ-**strongly convex**:

$$f(x) - \frac{\mu}{2}\|x\|^2 \text{ is convex}$$

Define $x^\star$ to be the unique minimizer of (1) and $f^\star := f(x^\star)$

Gradient descent (GD) method

GD scheme:

$$x_{k+1} = x_k - s\nabla f(x_k)$$

where $s > 0$ is the step size

The convergence rate of GD

- $\mu > 0$ **and** $s \in (0, 2/(L+\mu)]$:

$$\|x_k - x^\star\|^2 \leq \left(1 - s\frac{2\mu L}{\mu + L}\right)^k \|x_0 - x^\star\|^2$$

- $\mu = 0$ **and** $s \in (0, 1/L]$:

$$f(x_k) - f^\star \leq \frac{1}{2sk}\|x_0 - x^\star\|^2$$

Easy implementation, but converges slowly

Acceleration methods

- Heavy ball method (Polyak, 1964)

$$x_{k+1} = x_k - s\nabla f(x_k) + \alpha(x_k - x_{k-1})$$

  - α and s are constants
  - Local linear convergence for strongly convex functions
  - Global convergence fails for some choices of s

- Anderson acceleration methods (Anderson, 1965)

$$x_{k+1} = x_k - s_k\nabla f(x_k) - (X_k + s_k R_k)\Gamma_k$$

  - X_k, R_k are matrices from $x_k, \ldots, x_{k-m_k}$ and $\nabla f(x_k), \ldots, \nabla f(x_{k-m_k})$
  - Γ_k satisfies certain conditions
  - Accelerate fixed point iteration in computational physics, etc
  - The theoretical properties are underexplored

Nesterov accelerated gradient (NAG) method

- A seminar work proposed in Nesterov (1983)
- General NAG framework

$$\begin{cases} x_{k+1} = y_k - s\nabla f(y_k) \\ \beta_{k+1} = (t_{k+1} - 1)/t_{k+2} \\ y_{k+1} = x_{k+1} + \beta_{k+1}(x_{k+1} - x_k) \end{cases}$$

- $s \in (0, 1/L]$ is the step size and $\{t_k\}$ is a predefined sequence
- Easy implementation as GD
- Convergence speed **depends on the choice of extrapolation coefficients** $\{t_k\}$

Case I: $\mu > 0$ is known

$$t_k \equiv t^\star := \frac{\sqrt{L} + \sqrt{\mu}}{2\sqrt{\mu}} \quad \Longrightarrow \quad \beta_k \equiv \beta^\star := \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$$

If $s = 1/L$, we have

$$\begin{cases} x_{k+1} := y_k - \frac{1}{L}\nabla f(y_k) \\ y_{k+1} := x_{k+1} + \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}(x_{k+1} - x_k) \end{cases} \qquad \textbf{(NAG-sc)}$$

- Global R-linear convergence

$$f(x_k) - f^\star \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \left(f(x_0) - f^\star + \frac{\mu}{2}\|x_0 - x^\star\|^2\right).$$

- **Accurately estimating μ is challenging in practice**

Case II: $\mu = 0$ or $\mu > 0$ is unknown

Nesterov's rule

The sequence $\{t_k\}$ satisfies

$$t_1 = 1, \quad t_k \nearrow +\infty, \quad \text{and} \quad t_{k+1}^2 - t_{k+1} \le t_k^2, \quad \text{for} \quad k \ge 1$$

NAG-c: NAG method that satisfies Nesterov's rule with $s \in (0, 1/L]$

Two common choices of $\{t_k\}$ in NAG-c

1. $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2} = \sqrt{t_k^2 + \frac{1}{4}} + \frac{1}{2}$ (Nesterov, 1983)
2. $t_{k+1} = \frac{k+r}{r}$, with $r \ge 2$ (Lan et al., 2011; Tseng, 2008; Chambolle and Dossal, 2015; Attouch and Peypouquet, 2016; Su et al., 2016)

NAG-c has wide applications in image processing, machine learning, etc

Goal of this talk

Two questions related to NAG-c:

1. Whether NAG-c have global R-linear convergence for minimizing strongly convex problems?
   - Simplest case, but still unknown for more than 40 years
2. Can we the mathematical analysis of gradient restarted NAG-c over the NAG-c?
   - Classical acceleration techniques in extrapolation based methods, but establishing the theoretical advantages may be difficult

Contents

Motivation: numerical perspective

- NAG-c is faster than GD in convex setting ($O(1/k^2)$ v.s. $O(1/k)$)
- NAG-c has global R-linear convergence rather than Q-linear
- Fast intial convergence, slow linear asymptotic convergence

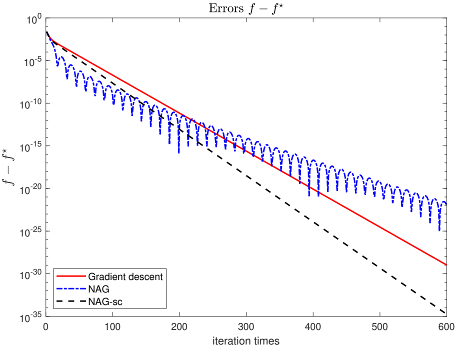


Figure: Numerical comparison between GD, NAG-c and NAG-sc

Motivation: ODE perspective

Setting $t_k = \frac{k+2}{2}$, NAG-c reduces to

$$\begin{aligned} x_{k+1} &= y_k - s\nabla f\left(y_k\right) \\ y_{k+1} &= x_{k+1} + \frac{k}{k+3}\left(x_{k+1} - x_k\right) \end{aligned} \tag{2}$$

Rewrite (2) as

$$\frac{x_{k+1} - x_k}{\sqrt{s}} = \frac{k-1}{k+r}\frac{x_k - x_{k-1}}{\sqrt{s}} - \sqrt{s}\nabla f\left(y_k\right)$$

Define $t = k\sqrt{s}$ and $x_k = X(t)$, then

$$\begin{aligned} &\dot{X}(t) + \frac{1}{2}\ddot{X}(t)\sqrt{s} + o(\sqrt{s}) \\ = &\left(1 - \frac{3\sqrt{s}}{t}\right)\left(\dot{X}(t) - \frac{1}{2}\ddot{X}(t)\sqrt{s} + o(\sqrt{s})\right) - \sqrt{s}\nabla f(X(t)) + o(\sqrt{s}) \end{aligned}$$

Ignoring $o(\sqrt{s})$ term, we get a low-resolution ODE (Su et al., 2016)

$$\begin{cases} \ddot{X}(t) + \frac{3}{t}\dot{X}(t) + \nabla f(X(t)) = 0 \\ X(0) = x_0, \ \dot{X}(0) = 0 \end{cases}$$

- Consistency result

$$\lim_{s \to 0} \max_{0 \le k \le T/\sqrt{s}} \left\| x_k - X(k\sqrt{s}) \right\| = 0$$

- If $f = \frac{1}{2}\langle x, \Lambda x \rangle$, it has

$$f(X(t)) - f^\star = O\left(\frac{\left\| x_0 - x^\star \right\|^2}{t^3 \sqrt{\min \lambda_i}} \right)$$

$$\limsup_{t \to \infty} t^3 \left(f(X(t)) - f^\star \right) \ge \frac{2 \left\| x_0 - x^\star \right\|^2}{\pi \sqrt{L}}$$

It rules out the possibility of linear convergence, and contradicts with our numerical observation

Current results

- Local linear convergence
  - Asymptotic linear convergence with rate $\sqrt{1 - \frac{\mu}{L}} + \epsilon$ (Tao et al., 2016; Liang et al., 2017)
  - Non-asymptotic linear convergence with rate $1 - \frac{(1-Ls)\mu s}{4}$ when $s < 1/L$ (Li et al., 2023)
- Global convergence:
  - Sublinear convergence $O(1/\mathbf{poly}(k))$ (Su et al., 2016; Aujol et al., 2023)
- Global linear convergence with additional constraints
  - NAG with fixed restarting (O'Donoghue and Candès, 2015)
  - NAG required that $\sup_k \beta_k < 1$ (Wen et al., 2017) .

Key result: $s < 1/L$

Define the Lyapunov sequence $\mathcal{E}_k$ as

$$s(t_{k+1}-1)t_{k+1}\left(f\left(x_k\right)-f^\star\right)+\frac{1}{2}\left\|(t_{k+1}-1)(y_k-x_k)+(y_k-x^\star)\right\|^2$$

Theorem (NAG-c: $s < 1/L$)

There exists a positive sequence $\{\rho_k\}$ such that for all $k \geq 1$,

$$\mathcal{E}_{k+1} \leq \rho_k \mathcal{E}_k, \quad \textit{and} \quad f\left(x_k\right)-f^\star \leq \frac{\prod_{i=1}^{k}\rho_i}{(t_{k+1}-1)t_{k+1}} \cdot \frac{\left\|x_0-x^\star\right\|^2}{2s},$$

with

$$\begin{cases} \bar{\rho}: & = \sup\limits_{k\geq 0} \rho_k \quad \leq 1-\frac{(1-Ls)\mu s}{1+\max\left\{\frac{\mu}{L},\frac{1}{8}\right\}} \quad \textit{(Global rate)} \\ \rho_\infty: & = \lim\limits_{k\to\infty} \rho_k \quad \leq 1-\frac{(1-Ls)\mu s}{1+\frac{\mu}{L}} \quad \textit{(Local rate)} \end{cases}$$

Comparison for convergence speed

K-step decreasing ratio

- GD: $(1-\mu s)^k$; NAG-sc: $(1-\sqrt{\mu}L)^k$; NAG-c: $\frac{\prod_{i=1}^{k}\rho_i}{(t_{k+1}-1)t_{k+1}}$

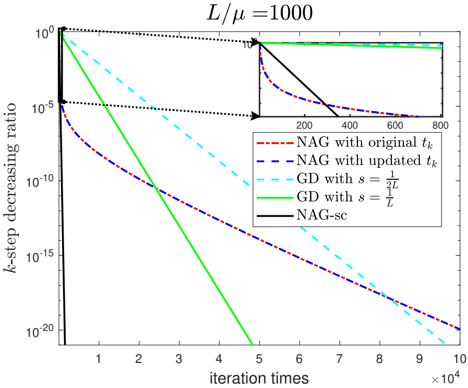


Figure: Numerical comparison with $L/\mu = 1000$

Sketch of the proof

- Descent property of $\{\mathcal{E}_k\}$:

$$\mathcal{E}_{k+1} - \mathcal{E}_k \leq - \frac{s^2 t_{k+1}^2 (1 - sL)}{2} \left\| \nabla f(y_k) \right\|^2 - \frac{\mu s (t_{k+1} - 1) t_{k+1}}{2} \left\| y_k - x_k \right\|^2 - \frac{\mu s t_{k+1}}{2} \left\| y_k - x^\star \right\|^2$$

- Boundedness of $\{\mathcal{E}_k\}$: for any $a, b > 0$, it has

$$\mathcal{E}_k \leq \frac{s(t_{k+1} - 1) t_{k+1} (1 + \mu/a)}{2\mu} \left\| \nabla f(y_k) \right\|^2 + \frac{1 + 1/b}{2} \left\| y_k - x^\star \right\|^2 + \left[\frac{(1+b)(t_{k+1} - 1)^2 + s(t_{k+1} - 1) t_{k+1} (a + L)}{2} \right] \left\| y_k - x_k \right\|^2$$

Similar bound can be proved for $\mathcal{E}_{k+1}$

We can prove that

$$\mathcal{E}_{k+1} \leq \rho_k \mathcal{E}_k \quad \text{with} \quad \rho_k := \left(1 - \frac{1}{\min\{\mathcal{C}_k, \mathcal{D}_k\}}\right)$$

- $\{\mathcal{C}_k\}$ is increasing from $\mathcal{C}_0 = 1/\mu s$ to

$$\lim_{k\to\infty} \mathcal{C}_k = \mathcal{C}_\infty := \frac{1+Ls}{\mu s} + \frac{(Ls)^2 + \sqrt{(Ls)^4 + 4(1-Ls)\mu s}}{2(1-Ls)\mu s}$$

- $\mathcal{D}_k \in (1 + \frac{1}{1-Ls}, \frac{3}{(1-sL)\mu s})$ and

$$\lim_{k\to\infty} \mathcal{D}_k = \mathcal{D}_\infty := \frac{1+\mu s}{\mu s} + \frac{\delta + \sqrt{\delta^2 + 4(1-Ls)\mu s}}{2(1-Ls)\mu s} < \mathcal{C}_\infty$$

  where $\delta := (L-\mu)s + L\mu s^2$

This can easily obtain $\bar{\rho}$ and ρ_∞

Key result: $s = 1/L$

Define the Lyapunov sequence as

$$\mathcal{E}_k := \lambda(f(x_k) - f^\star) + \frac{1}{2}\left\|x_k - x_{k-1}\right\|^2, \quad \forall k \geq 0$$

where λ is a number depends on L and μ

Theorem (NAG-c: $s = 1/L$)

There exists a positive number ρ such that for all $k \geq 1$,

$$\mathcal{E}_{k+1} \leq \rho \mathcal{E}_k, \quad \text{and} \quad f(x_k) - f^\star \leq \rho^k (f(x_0) - f^\star)$$

with $\rho = 0$ if $\mu = L$, and

$$0 < \rho < \frac{4L^2 - 3L\mu}{4L^2 - 3L\mu + \mu^2} < 1 \quad \text{if } \mu < L$$

Remarks

- These results can be extended for accelerated proximal gradient methods (Tseng, 2008; Beck and Teboulle, 2009)
- Tightness of this bound is unknown
- Does there exists an ODE model consistent with the NAG-c method?
  - The high-resolution ODE model (Shi et al., 2022)

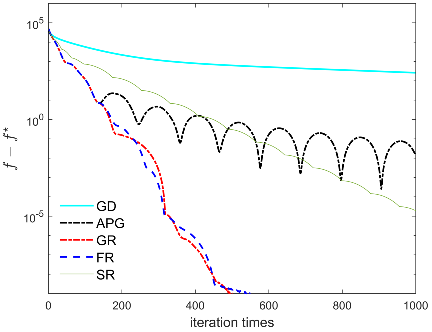
$$\ddot{X}(t) + \frac{3}{t}\dot{X}(t) + \sqrt{s}\nabla^2 f(X(t))\dot{X}(t) + \left(1 + \frac{3\sqrt{s}}{2t}\right)\nabla f(X(t)) = 0$$

  - Distinguish the heavy ball method and NAG methods
  - Provable locally linear convergence for $s < 1/L$
  - Global linear convergence is unknown

Contents

Adaptive restart schemes

- Motivation: avoid the oscillation phenomenon of NAG-c
- Restart: reset $\beta_k = 0$ if some conditions (O'Donoghue and Candès, 2015; Beck and Teboulle, 2009; Su et al., 2016) are met
  - **Gradient restart:** $\langle x_k - x_{k-1}, y_{k-1} - x_k \rangle > 0$
  - Function value restart: $f(x_k) < f(x_{k=1})$
  - Speed restart: $|x_k - x_{k-1}| < |x_{k-1} - x_{k-2}|$



Gradient restarted NAG-c

Recall the low-resolution ODE:

$$\begin{cases} \ddot{X}(t) + \frac{3}{t}\dot{X}(t) + \nabla f(X(t)) = 0 \\ X(0) = x_0, \ \dot{X}(0) = 0 \end{cases} \qquad \text{(NAG-ODE)}$$

Gradient restarted scheme: reset $t = 0$ when

$$\langle \nabla f(X(t)), \dot{X}(t) \rangle \geq 0$$

- If $f(x) = \frac{1}{2}\langle x, \Lambda x \rangle$ where $\Lambda \succ 0$, NAG-c has sublinear Convergence

$$f(X(t)) - f^* \geq O(1/t^3), \qquad \text{(Optimal rate)}$$

- Whether the gradient restart scheme has global linear convergence for strongly convex problems is open (Su et al., 2016)

ODE for gradient restarted NAG-c

Define the gradient restart time:

$$T^{\mathrm{gr}}(x_0; f) = \sup\left\{t > 0 \mid \langle \nabla f(X(u)), \dot{X}(u)\rangle < 0, \forall u \in (0, t)\right\}.$$

- Let $E_0 = 0$ and $r_0 = x_0$, and

$$E_{i+1} = T^{\mathrm{gr}}(r_i; f) \quad \text{and} \quad r_{i+1} = Y_{i+1}(E_{i+1}),$$

  where $Y_{i+1}(t)$ solves NAG-ODE with $x_0 = r_i$.
- The gradient restarted NAG-ODE:

$$\begin{cases} \ddot{X}(t) + \frac{3}{t-\tau_i}\dot{X}(t) + \nabla f(X(t)) = 0, \quad \text{for} \quad t \in (\tau_i, \tau_{i+1}], \\ X(\tau_i) = r_i, \quad \dot{X}(\tau_i) = 0, \end{cases} \tag{3}$$

  where $\tau_i := \sum_{j=0}^{i} E_j$, $i \geq 0$

Global R-linear convergence

Assumption (uniform unpper bound)

Given $f \in \mathcal{S}_{\mu,L}$, there exists $T > 0$ such that $T^{\mathrm{gr}}(x_0; f) \leq T$ for all $x_0 \in \mathbb{R}^n$.

Theorem

Assume $f \in \mathcal{S}_{\mu,L}$ *and suppose the above assumption holds, then there exist positive constants* $c_1 > 0$ *and* $c_2 \in (0,1)$*, which only depend on* L*,* μ *and* T*, such that*

$$f(X^{\mathrm{gr}}(t)) - f^\star \leq \frac{c_1 L \|x_0 - x^\star\|^2}{2} e^{-c_2 t}$$

where $X^{\mathrm{gr}}(t)$ *is the solution of* (3)

Validation of the assumption

Consider $f(x) = \frac{1}{2}\langle x, \Lambda x\rangle$, $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$ and $\lambda_1 \geq \cdots \geq \lambda_n > 0$

- Define

$$H(t) = \langle \nabla f(X(t), \dot{X}(t)\rangle = \sum_{i=1}^{n} \lambda_i X_i(t) \dot{X}_i(t)$$

  where X_i satisfies $\ddot{X}_i + \frac{3}{t}\dot{X}_i + \lambda X_i = 0$ with $X_i(0) = x_{0,i}$

- Validating the uniform upper bound assumption is equivalent to

  $\exists \mathbf{T} > \mathbf{0}$ independent with $\mathbf{x_0}$ and $\mathbf{t_{x_0}} \in (\mathbf{0}, \mathbf{T}]$ such that $\mathbf{H}(\mathbf{t_{x_0}}) \geq \mathbf{0}$

- X_i has the form

$$X_i(t) = \frac{2x_{0,i}}{t\sqrt{\lambda_i}} J_1(\sqrt{\lambda_i} t)$$

  where J_1 is the Bessel function of the first kind with order 1

$$H(t) = \sum_{i=1}^{n} H_i(t) \quad \text{with} \quad H_i(t) = -\frac{4\sqrt{\lambda_i} x_{0,i}^2}{t^2} J_1(\sqrt{\lambda_i} t) J_2(\sqrt{\lambda_i} t)$$

Define $G(u) = \pi u J_1(u) J_2(u)$, then $H_i(t) = -\frac{4x_{0,i}^2}{\pi t^3} G(\sqrt{\lambda_i} t)$

Two key lemmas

- **Asymptotic behavior of G:**

$$|G(u) - \cos(2u)| \leq \epsilon, \quad \forall u > T_\epsilon$$

  Leads to oscillation phenomenon when t is large

- **Second form Kronecker's theorem:** Let $1, \alpha_1, \ldots, \alpha_s \in \mathbb{R}$ be linearly independent over rationals, then the set

$$\{\mathbf{frac}(\nu\boldsymbol{\alpha}) | \nu \in \mathbb{N}\} = \{(\mathbf{frac}(\nu\alpha_1), \ldots, \mathbf{frac}(\nu\alpha_s)) \, | \nu \in \mathbb{N}\} \subset \mathbb{R}^s$$

  is dense in $[0, 1]^s$
  critical for the high dimensional case when $\sqrt{\lambda_i}/\sqrt{\lambda_j} \notin \mathbb{Q}$

Consider the quadratic case

$$f(x) = \frac{1}{2}\langle x, Ax\rangle + \langle x, b\rangle, \tag{4}$$

where $A \in \mathbb{R}^{n\times n}$ is a symmetric and positive definite matrix and $b \in \mathbb{R}^n$.

Theorem

Let f be defined in (4), and $X^{\mathrm{gr}}(t)$ is the solution of the gradient restarted NAG-ODE. Then, $f(X^{\mathrm{gr}}(t))$ converges to $f^\star$ at a globally R-linear rate

- This result partially solves the open problem
- But technical difficult remains when extending this proof to the general strongly convex case

Contents

Non-smooth case

Model

$$\min_{x} \; F(x) = f(x) + g(x)$$

where f: L-smooth and μ-strongly convex; g: convex

The **APG-c** (Tseng, 2008; Beck and Teboulle, 2009) has

$$\begin{cases} x_{k+1} := \mathbf{prox}_{sg}(y_k - s\nabla f(y_k)) \\ \beta_{k+1} := (t_{k+1} - 1)/t_{k+2} \text{ with } t_{k+2} \text{ satisfies Nesterov's Rule} \\ y_{k+1} := x_{k+1} + \beta_{k+1}(x_{k+1} - x_k) \end{cases}$$

where the proximal mapping $\mathbf{prox}_g : \mathbb{R}^n \to \mathbb{R}^n$ of g is defined by

$$\mathbf{prox}_g(y) := \arg\min_{x \in \mathbb{R}^n} \left\{ g(x) + \frac{1}{2}\|x - y\|_2^2 \right\}, \quad \forall y \in \mathbb{R}^n$$

Rate comparison

- GR+APG: the gradient restarted APG method;
- UBC: the condition that the restart intervals are uniformly bounded;
- †: we assume that t_k satisfies the common choices.

Objective function	$f+g$	$f+g$	
Algorithm	APG-c	GR+APG-c	
		Original	+UBC
$\|x^k - x^*\|$	$O(\bar{\rho}^{k-1}/k)^\dagger$	$O(\bar{\rho}^k)$	$O(\hat{\rho}^k), \hat{\rho} < \bar{\rho}$

Table: Rate comparison between the APG-c and gradient restarted APG if $s < 1/L$.

Remark: optimal restart interval depends on L, μ and $f^\star$ (Aujol et al., 2023)

Multi-step extrapolation based methods

- Define $r_k = -\nabla f(X_k)$, X_k and R_k to be

$$X_k = [\Delta x_{k-m_k}, \Delta x_{k-m_k+1}, \cdots, \Delta x_{k-1}]$$
$$R_k = [\Delta r_{k-m_k}, \Delta r_{k-m_k+1}, \cdots, \Delta r_{k-1}]$$

- Anderson acceleration scheme

$$x_{k+1} = x_k + s_k r_k - (X_k + s_k R_k)\Gamma_k$$

  - Type-I: $r_k - R_k\Gamma_k \perp \mathrm{Range}(X_k)$
  - Type-II: $r_k - R_k\Gamma_k \perp \mathrm{Range}(R_k)$

- Wide applications in computational physics, etc

Restart: reset $X_k = [\ ]$ and $R_k = [\ ]$ if

- $m_k \leq n$
- $|v_k^\top q_k| \geq \tau |v_{k-m_k+1}^\top q_{k-m_k+1}|, \quad \tau \in (0,1)$
- $\|r_k\|_2 \leq \eta \|r_{k-m_k}\|_2, \quad \eta \geq 1$

Local convergence results for restarted AM

- Type I AM:

$$\theta_k \sqrt{1+\gamma_k^2 \kappa_k^2} \min_{\substack{p \in \mathcal{P}_{m_k} \\ p(0)=1}} \|p(A)(x_{k-m_k} - x^*)\|_2 + \hat{\kappa}\mathcal{O}(\|x_{k-m_k} - x^*\|_2^2),$$

where $\gamma_k \leq L$, $\kappa_k \leq 1/\mu$, $\theta_k = \|I - \beta_k A\|_2$ and $A = \nabla^2 f(x^*)$

- Type II AM:

$$\theta_k \min_{\substack{p \in \mathcal{P}_{m_k} \\ p(0)=1}} \|p(A) r_{k-m_k}\|_2 + \hat{\kappa}\mathcal{O}(\|x_{k-m_k} - x^*\|_2^2).$$

A remark

Important questions mentioned by recent review (100 pages)

- The adaptive choice of m
- Numerical and model improvements on extrapolation coefficients
- The convergence analysis when G is not contractive/nonsmooth
- The effects of restart technique

Numer Algor (2019) 80:135–234
https://doi.org/10.1007/s11075-018-0549-4

CrossMark

ORIGINAL PAPER

Comments on "Anderson Acceleration, Mixing and Extrapolation"

Donald G. M. Anderson[1]

Summary

- Global R-linear convergence of NAG-c in strongly convex setting
  - Needs an ODE that consistent with the discretized algorithm
- Mathematical analysis for the gradient restarted NAG-C
  - Fully solving the open problem in Su et al. (2016) requires new tools
- Local convergence rate of restarted Anderson acceleration
  - Theoretical analysis for multi-step extrapolation methods is under explored, such as limited-memory Anderson acceleration, restarted Halpern iteration

References:

1. **B.**, L. Chen, J. Li. The Global R-linear Convergence of Nesterov's Accelerated Gradient Method with Unknown Strongly Convex Parameter, *ArXiv:2308.14080*, 2023

2. **B.**, L. Chen, J. Li, Z. Shen. Accelerated Gradient Methods with Gradient Restart: Global Linear Convergence, *ArXiv: 2401.07672*, 2024

3. F. Wei, **B.**, Y. Liu, G. Yang. Convergence Analysis for Restarted Anderson Mixing and Beyong, *ArXiv: 2307.02062*, 2023

Thank you!

Reference I

Hédy Attouch and J. Peypouquet. The rate of convergence of nesterov's accelerated forward-backward method is actually faster than $1/k^2$. *SIAM Journal on Optimization*, 26(3):1824–1834, 2016.

J-F Aujol, Ch Dossal, and Aude Rondepierre. FISTA is an automatic geometrically optimized algorithm for strongly convex functions. *Mathematical Programming*, pages 1–43, 2023.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

A Chambolle and Ch Dossal. On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". *Journal of Optimization Theory and Applications*, 166(3):968–982, 2015.

Guanghui Lan, Zhaosong Lu, and Renato D. C. Monteiro. Primal-dual first-order methods with $\mathcal{O}(1/\epsilon)$ iteration-complexity for cone programming. *Mathematical Programming*, 126:1–29, 2011.

Bowen Li, Bin Shi, and Ya xiang Yuan. Linear convergence of Nesterov-1983 with the strong convexity. 2023.

Reference II

Jingwei Liang, Jalal Fadili, and Gabriel Peyré. Activity identification and local linear convergence of forward-backward-type methods. *SIAM Journal on Optimization*, 27(1): 408–437, 2017.

Yurii Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. In *Soviet Mathematics Doklady*, pages 372–376, 1983.

Brendan O'Donoghue and Emmanuel J. Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15:715–732, 2015.

Bin Shi, Simon S Du, Michael I Jordan, and Weijie J Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, 195 (1):79–148, 2022.

Weijie Su, Stephen P. Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17:153:1–153:43, 2016.

Shaozhe Tao, Daniel Boley, and Shuzhong Zhang. Local linear convergence of ISTA and FISTA on the LASSO problem. *SIAM Journal on Optimization*, 26(1):313–336, 2016.

Reference III

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. Online, 2008. Preprint at https://www.mit.edu/~dimitrib/PTseng/papers/apgm.pdf.

Bo Wen, Xiaojun Chen, and Ting Kei Pong. Linear convergence of proximal gradient algorithm with extrapolation for a class of nonconvex nonsmooth minimization problems. *SIAM Journal on Optimization*, 27(1):124–145, 2017.