On the Acceleration of First-order Methods

and a New One

Jingwei LIANG

Institute of Natural Sciences, Shanghai Jiao Tong University

Joint work with: **Clarice Poon (Bath UK)**

Outline

Non-smooth optimization

Example - Constrained non-smooth optimization

$$\min_{x \in \mathbb{R}^n} \; F(x) + \sum\nolimits_{i=1}^{r} R_i(K_i x),$$

where

- F: smooth differentiable with L-Lipschitz continuous gradient.
- R_i: proper and lower semi-continuous.
- K_i: bounded linear mapping.

Applications: signal/image processing, inverse problems, machine learning, data science, statistics...

Challenges: non-smooth, (non-convex), composite, high dimension...

Hope: well structured.

Example: blind image deconvolution

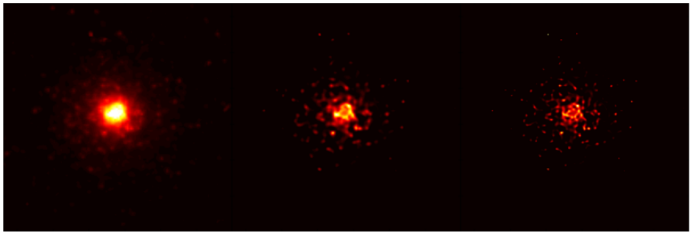


Figure: NGC224 by Hubble Space Telescope from Wikipedia.

Forward model

$$A = y \odot x + \varepsilon.$$

Example: blind image deconvolution

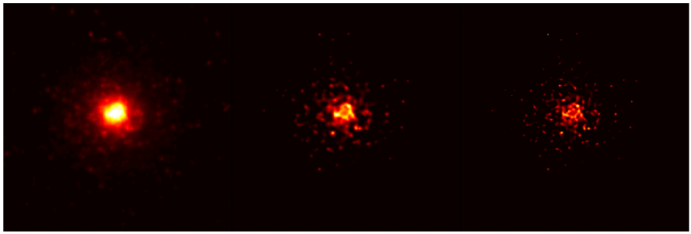


Figure: NGC224 by Hubble Space Telescope from Wikipedia.

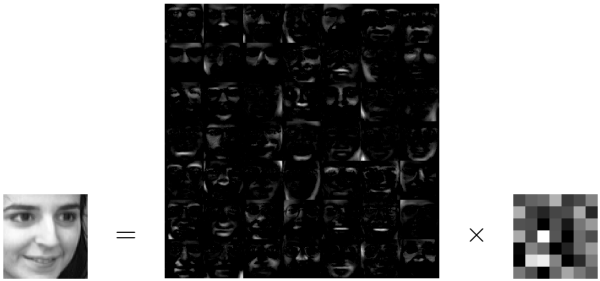
Forward model

$$A = y \odot x + \varepsilon.$$

Example - blind image deconvolution

$$\min_{x \in \mathbb{R}^{p \times r}, y \in \mathbb{R}^{s \times s}} \frac{1}{2}\|A - y \odot x\|^2 + \lambda R(x) \quad \text{s.t.} \quad 0 \preceq x \preceq 1,\ 0 \preceq y \preceq 1,\ \|y\|_1 = 1.$$

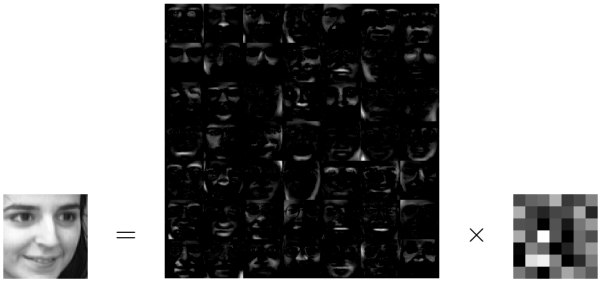
Example: sparse non-negative matrix factorization



Decomposition model

$$A = yx + \varepsilon.$$

Example: sparse non-negative matrix factorization



Decomposition model

$$A = yx + \varepsilon.$$

Example - sparse NMF

$$\min_{x \in \mathbb{R}^{p \times r}, y \in \mathbb{R}^{r \times q}} \tfrac{1}{2}\|A - xy\|^2 \quad \text{s.t.} \quad x, y \geq 0, \ \ \|x_i\|_0 \leq \kappa_x, \ i = 1, ..., r.$$

First-order methods: two basic ingredients

Gradient descent [Cauchy '1847]

$$\min_x F(x)$$

where F is convex smooth differentiable with ∇F being L-Lipschitz.

First-order methods: two basic ingredients

Gradient descent [Cauchy '1847]

$$\min_x F(x)$$

where F is convex smooth differentiable with ∇F being L-Lipschitz.

EXplicit Euler scheme:

$$x_{k+1} = x_k - \gamma_k \nabla F(x_k), \quad \gamma_k \in]0, 2/L[.$$

First-order methods: two basic ingredients

Gradient descent [Cauchy '1847]

$$\min_x F(x)$$

where F is convex smooth differentiable with ∇F being L-Lipschitz.

EXplicit Euler scheme:

$$x_{k+1} = x_k - \gamma_k \nabla F(x_k), \quad \gamma_k \in]0, 2/L[.$$

Proximal point algorithm [Rockafellar '76]

$$\min_x R(x)$$

with R being proper closed convex.

First-order methods: two basic ingredients

Gradient descent [Cauchy '1847]

$$\min_x F(x)$$

where F is convex smooth differentiable with ∇F being L-Lipschitz.
EXplicit Euler scheme:

$$x_{k+1} = x_k - \gamma_k \nabla F(x_k), \quad \gamma_k \in]0, 2/L[.$$

Proximal point algorithm [Rockafellar '76]

$$\min_x R(x)$$

with R being proper closed convex. Define "proximity operator" by

$$\mathrm{prox}_{\gamma R}(v) \overset{\text{def}}{=} \mathrm{argmin}_x \ \gamma R(x) + \frac{1}{2}\|x - v\|^2.$$

First-order methods: two basic ingredients

Gradient descent [Cauchy '1847]

$$\min_x F(x)$$

where F is convex smooth differentiable with ∇F being L-Lipschitz.

EXplicit Euler scheme:

$$x_{k+1} = x_k - \gamma_k \nabla F(x_k), \quad \gamma_k \in]0, 2/L[.$$

Proximal point algorithm [Rockafellar '76]

$$\min_x R(x)$$

with R being proper closed convex. Define "proximity operator" by

$$\mathrm{prox}_{\gamma R}(v) \overset{\text{def}}{=} \mathrm{argmin}_x \ \gamma R(x) + \tfrac{1}{2}\|x - v\|^2.$$

IMplicit Euler scheme:

$$x_{k+1} = \mathrm{prox}_{\gamma_k R}(x_k), \quad \gamma_k > 0.$$

First-order methods

A rich class of first-order methods

$F + R$	Forward–Backward splitting [Lions & Mercier '79]
$R_1 + R_2$	Douglas–Rachford splitting [Douglas & Rachford '56; Lions & Mercier '79]
	ADMM [Glowinski & Marrocco '75; Gabay & Mercier '76]...
$F + R(K\cdot)$	Primal–Dual splitting methods [Arrow, Hurwicz & Uzawa '58; Esser, Zhang & Chan '10; Chambolle & Pock '11]
$F + \sum_i R_i$	Generalized Forward–Backward splitting [Raguet, Fadili & Peyré '13]
–	...

`Origins from numerical PDE back to 1950s, now ubiquitous in signal/image processing, inverse problems, data science, statistics, machine learning...`

First-order methods

A rich class of first-order methods

$F + R$	Forward–Backward splitting [Lions & Mercier '79]
$R_1 + R_2$	Douglas–Rachford splitting [Douglas & Rachford '56; Lions & Mercier '79]
	ADMM [Glowinski & Marrocco '75; Gabay & Mercier '76]...
$F + R(K\cdot)$	Primal–Dual splitting methods [Arrow, Hurwicz & Uzawa '58; Esser, Zhang & Chan '10; Chambolle & Pock '11]
$F + \sum_i R_i$	Generalized Forward–Backward splitting [Raguet, Fadili & Peyré '13]
–	...

Fixed-point formulation

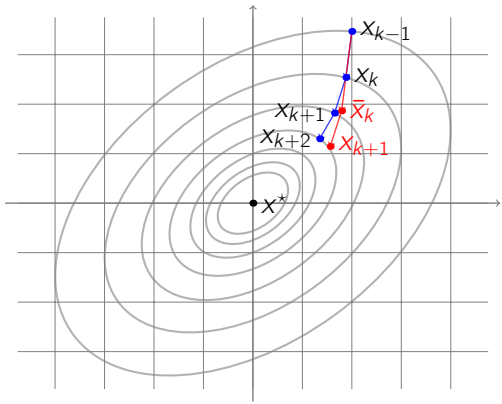
Let $\mathcal{F} : \mathcal{H} \to \mathcal{H}$ be non-expansive with $\operatorname{fix}(\mathcal{F}) \stackrel{\text{def}}{=} \{z \in \mathcal{H} \mid z = \mathcal{F}(z)\} \neq \emptyset$,

$$z_{k+1} = \mathcal{F}(z_k).$$

Accelerations: two approaches

Definition - Inertial technique [Polyak '64, Nesterov '83, Beck & Teboulle '09]

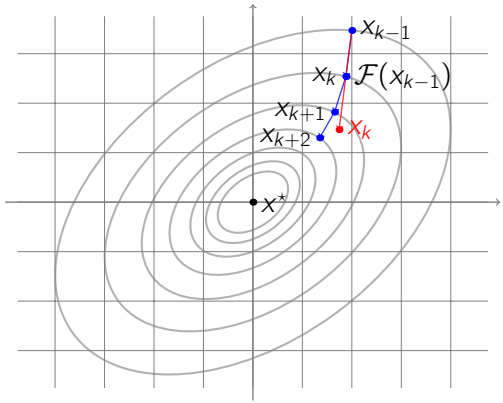
$$\begin{cases} \bar{x}_k = x_k + a_k(x_k - x_{k-1}), \\ x_{k+1} = \mathcal{F}(\bar{x}_k). \end{cases}$$



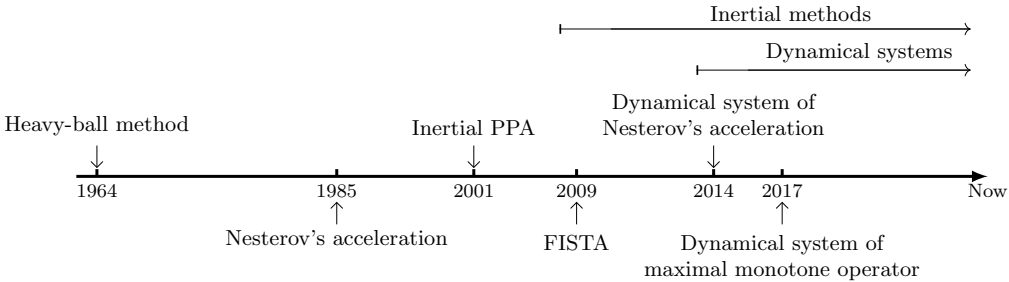
Accelerations: two approaches

Definition - Successive over-relaxation [Richardson '1911, Young '50]

$$x_{k+1} = (1-\lambda_k)x_k + \lambda_k \mathcal{F}(x_k) \xrightarrow{a_k=\lambda_k-1} \begin{cases} \bar{x}_k = x_k + a_k(x_k - \bar{x}_{k-1}), \\ x_{k+1} = \mathcal{F}(\bar{x}_k). \end{cases}$$



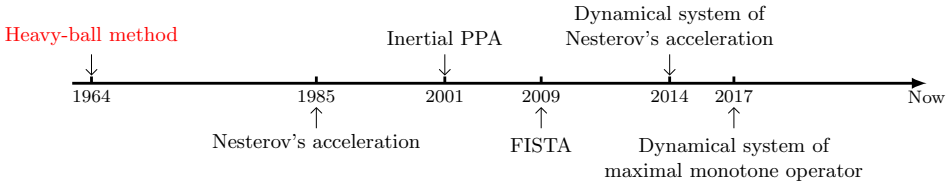
A brief history of inertial acceleration



A brief timeline of inertial acceleration.

A glimpse of references: [Polyak '64], [Polyak '87], [Nesterov '07], [Nesterov '83], [Nesterov '04], [Alvarez '00], [Attouch, Goudou & Redont '00], [Alvarez & Attouch '01], [H Attouch, Peypouquet & Redont '14], [Attouch & Peypouquet '16], [Attouch & Peypouquet '16], [Attouch & Peypouquet '19], [Adly & Attouch '20], [Moudafi & Oliny '03], [Maingé '08], [Beck & Teboulle '09], [O'Donoghue '12], [Su, Boyd & Candès '14], [Su, Boyd & Candès '16], [Chambolle & Dossal '15], [Aujol & Dossal '15], [Lorenz & Pock '14], [Boţ, Csetnek & Hendrich '15], [Boţ & Csetnek '16] [Liang, Fadili and Gabriél '17], [Apidopoulos, Aujol & Dossal '20], [França, Robinson and Vidal '18a], [França, Robinson and Vidal '18b], [Calatroni & Chambolle '19], [Chambolle & Tovey '21] and many others...

A brief history of inertial acceleration



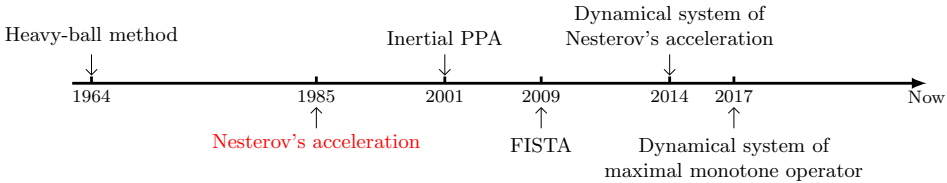
[Polyak '64] Consider minimizing $F \in C^1_L(\mathbb{R}^n)$

$$\begin{cases} \bar{x}_k = x_k + a_k(x_k - x_{k-1}), \\ x_{k+1} = \bar{x}_k - \gamma \nabla F(x_k). \end{cases}$$

- If $F \in C^2_{L,\mu}(\mathbb{R}^n)$, the optimal rate can be achieved $\rho^\star = \frac{1-\sqrt{\mu/L}}{1+\sqrt{\mu/L}}$.
- Heavy-ball dynamics with friction:

$$\ddot{x}(t) + \gamma \dot{x}(t) = -\nabla F(x(t)).$$

A brief history of inertial acceleration

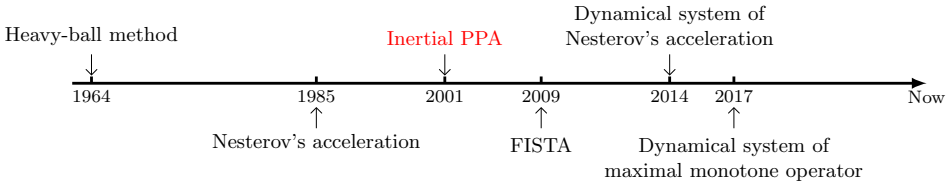


[Nesterov '83] Consider minimizing $F \in C^1_L(\mathbb{R}^n)$: $\gamma \leq 1/L$

$$\begin{cases} \bar{x}_k = x_k + \dfrac{k-1}{k+2}(x_k - x_{k-1}), \\ x_{k+1} = \bar{x}_k - \gamma \nabla F(\bar{x}_k). \end{cases}$$

- Convex case: $o(1/k) \to o(1/k^2)$.
- If $F \in C^1_{L,\mu}(\mathbb{R}^n)$, $1 - \frac{\mu}{L} \to 1 - \sqrt{\frac{\mu}{L}}$.

A brief history of inertial acceleration



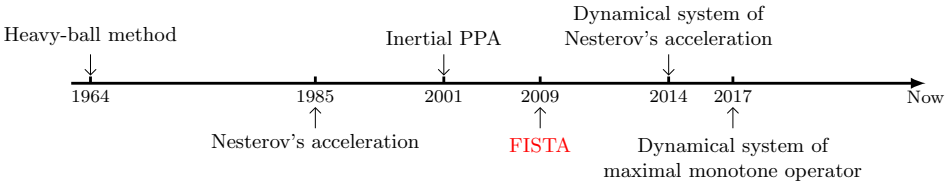
[Alvarez & Attouch '01] Consider maximal monotone inclusion problem $0 \in A(x)$

$$\ddot{x}(t) + \gamma \dot{x}(t) \in -A(x(t)).$$

- No co-coercivity.
- Guaranteed convergence for $a_k < \frac{1}{3}$ with

$$x_{k+1} = (\mathrm{Id} + \lambda_k A)^{-1}(x_k + a_k(x_k - x_{k-1})).$$

A brief history of inertial acceleration

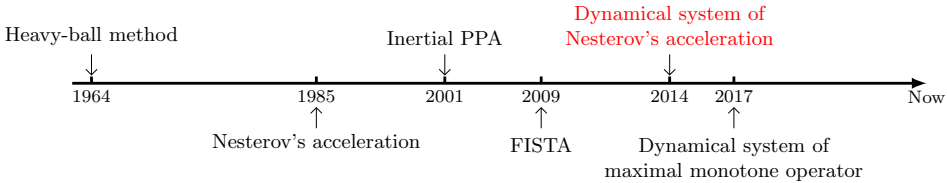


[Beck & Teboulle '09][Chambolle & Dossal '15] $F \in C_L^1(\mathbb{R}^n)$, $R \in \Gamma_0(\mathbb{R}^n)$: $\gamma \in]0, 1/L]$, $d > 2$

$$\bar{x}_k = x_k + \tfrac{k-1}{k+d}(x_k - x_{k-1}),$$
$$x_{k+1} = \mathrm{prox}_{\gamma R}(\bar{x}_k - \gamma \nabla F(\bar{x}_k)).$$

- Rate of convergence: sequence $o(1/\sqrt{k}) \to o(1/k)$, objective $o(1/k) \to o(1/k^2)$.
- Convergence of sequence.

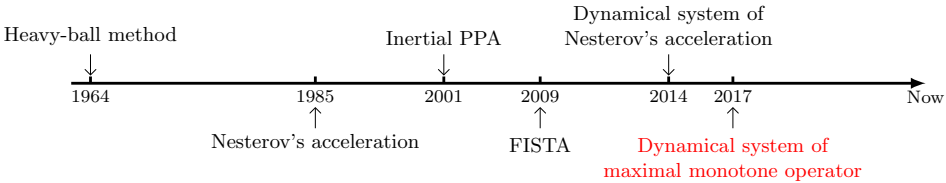
A brief history of inertial acceleration



[Su, Boyd & Candès '14] A differential equation for Nesterov's acceleration

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) = -\nabla F(x(t)).$$

A brief history of inertial acceleration



[Attouch & Peypouquet '19] Yosida's regularization

$$A_\lambda = \frac{1}{\lambda}\big(\mathrm{Id} - (\mathrm{Id} + \lambda A)^{-1}\big).$$

Damped inertial dynamics

$$\ddot{x}(t) + \gamma(t)\dot{x}(t) = -A_{\lambda(t)}(x(t)).$$

- $\gamma(t) = \frac{\alpha}{t}$ and $\lambda(t)\gamma^2(t) > 1$.
- Sequence rate of convergence: $o(1/\sqrt{k}) \to o(1/k)$.

Are we blessed with guaranteed acceleration?

Problem -

$$\min_{x\in\mathbb{R}^n} J(x) + R(x).$$

Let $\gamma > 0$

$$\mathcal{F}_{\mathrm{DR}} \overset{\text{def}}{=} \frac{1}{2}(\mathrm{Id} + (2\mathrm{prox}_{\gamma R} - \mathrm{Id})(2\mathrm{prox}_{\gamma J} - \mathrm{Id})).$$

Are we blessed with guaranteed acceleration?

Problem -

$$\min_{x\in\mathbb{R}^n} J(x) + R(x).$$

Let $\gamma > 0$

$$\mathcal{F}_{\text{DR}} \stackrel{\text{def}}{=} \frac{1}{2}(\text{Id} + (2\text{prox}_{\gamma R} - \text{Id})(2\text{prox}_{\gamma J} - \text{Id})).$$

Douglas–Rachford splitting [Douglas & Rachford '56]

$$z_{k+1} = \mathcal{F}_{\text{DR}}(z_k),$$

- Sequence $o(1/\sqrt{k})$, objective **NA**.

Are we blessed with guaranteed acceleration?

Problem -

$$\min_{x\in\mathbb{R}^n} J(x) + R(x).$$

Let $\gamma > 0$

$$\mathcal{F}_{\mathrm{DR}} \stackrel{\text{def}}{=} \frac{1}{2}(\mathrm{Id} + (2\mathrm{prox}_{\gamma R} - \mathrm{Id})(2\mathrm{prox}_{\gamma J} - \mathrm{Id})).$$

Inertial Douglas–Rachford [Boţ, Csetnek & Hendrich '15]

$$\begin{aligned} \bar{z}_k &= z_k + a_k(z_k - z_{k-1}), \\ z_{k+1} &= \mathcal{F}_{\mathrm{DR}}(\bar{z}_k). \end{aligned}$$

- No rates available, **may fail to provide acceleration.**

Are we blessed with guaranteed acceleration?

Problem -

$$\min_{x\in\mathbb{R}^n} J(x) + R(x).$$

Let $\gamma > 0$

$$\mathcal{F}_{\mathrm{DR}} \overset{\text{def}}{=} \frac{1}{2}(\mathrm{Id} + (2\mathrm{prox}_{\gamma R} - \mathrm{Id})(2\mathrm{prox}_{\gamma J} - \mathrm{Id})).$$

Regularized inertial Douglas–Rachford: let $\mathcal{J}_A = (\mathrm{Id} + A)^{-1} = \mathcal{F}_{\mathrm{DR}}$, $\alpha > 2$ and $\beta > 0$

$$\begin{aligned} y_k &= x_k + \frac{k-1}{k+\alpha}(x_k - x_{k-1}), \\ x_{k+1} &= \mathcal{J}_{\beta A_{\gamma_k}}(y_k). \end{aligned}$$

Let $\gamma_k = (1+\epsilon)\frac{\beta}{\alpha^2}k^2$ with $\epsilon > \frac{2}{\alpha-2}$,

- $\|x_k - x_{k-1}\| = o(1/k)$, $x_k \to x^\star \in \mathrm{zer}(A)$.

Are we blessed with guaranteed acceleration?

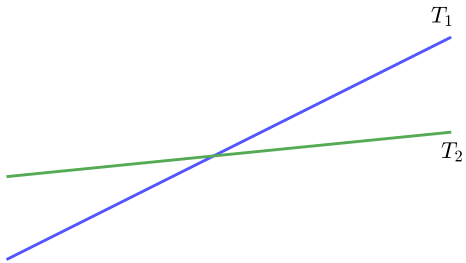
Problem - Feasibility problem in $\mathbb{R}^2$

Let $T_1, T_2 \subset \mathbb{R}^2$ be two subspaces such that $T_1 \cap T_2 \neq \emptyset$,

$$\text{Find } x \in \mathbb{R}^2 \text{ such that } x \in T_1 \cap T_2.$$

Define

$$\mathcal{F}_{\mathrm{DR}} \overset{\text{def}}{=} \frac{1}{2}(\mathrm{Id} + (2\mathcal{P}_{T_1} - \mathrm{Id})(2\mathcal{P}_{T_2} - \mathrm{Id})).$$



Are we blessed with guaranteed acceleration?

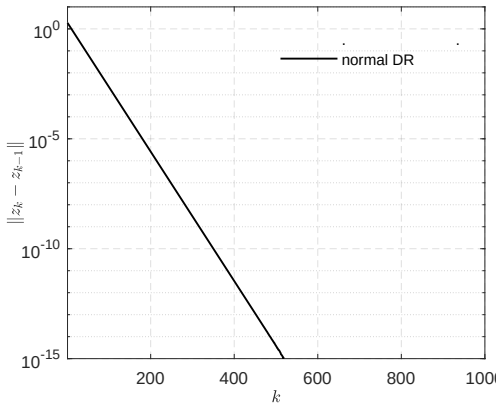
Problem - Feasibility problem in $\mathbb{R}^2$

Let $T_1, T_2 \subset \mathbb{R}^2$ be two subspaces such that $T_1 \cap T_2 \neq \emptyset$,

Find $x \in \mathbb{R}^2$ such that $x \in T_1 \cap T_2$.

Douglas–Rachford:

$$\bar{z}_k = z_k,$$
$$z_{k+1} = \mathcal{F}_{\mathrm{DR}}(\bar{z}_k).$$



Are we blessed with guaranteed acceleration?

Problem - Feasibility problem in $\mathbb{R}^2$

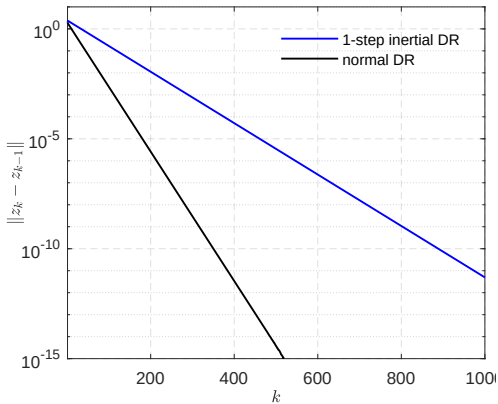
Let $T_1, T_2 \subset \mathbb{R}^2$ be two subspaces such that $T_1 \cap T_2 \neq \emptyset$,

Find $x \in \mathbb{R}^2$ such that $x \in T_1 \cap T_2$.

Inertial Douglas–Rachford:

$$\bar{z}_k = z_k + a(z_k - z_{k-1}),$$
$$z_{k+1} = \mathcal{F}_{\mathrm{DR}}(\bar{z}_k).$$

- 1-step inertial: $a = 0.3$.



Are we blessed with guaranteed acceleration?

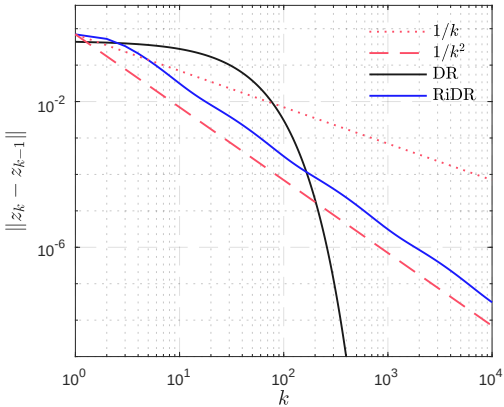
Problem - Feasibility problem in $\mathbb{R}^2$

Let $T_1, T_2 \subset \mathbb{R}^2$ be two subspaces such that $T_1 \cap T_2 \neq \emptyset$,

Find $x \in \mathbb{R}^2$ such that $x \in T_1 \cap T_2$.

Regularized inertial Douglas–Rachford:

$$y_k = x_k + \frac{k-1}{k+\alpha}(x_k - x_{k-1}),$$
$$x_{k+1} = \mathcal{J}_{\beta A_{\gamma_k}}(y_k).$$



Are we blessed with guaranteed acceleration?

Consider minimizing

$$\min_{x\in\mathbb{R}^n} R(x) + J(z) \quad \text{s.t.} \quad z = Ax.$$

Proposition - Dynamics of ADMM [França et al '18a]

Let $V(x) = R(x) + J(Ax)$,

$$(A^TA)\big(\ddot{x}(t) + \tfrac{\alpha}{t}\dot{x}(t)\big) = -\nabla V(x(t)).$$

- Smoothness are required by both R and J.
- Improve the objective rate of convergence from $O(1/t)$ to $O(1/t^2)$.
- With non-smooth objective, Yosida regularization can be applied [França et al '18b].
- Not result available in discrete setting...

Are we blessed with guaranteed acceleration?

Some remarks

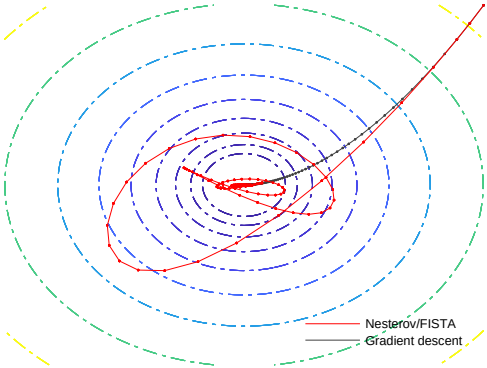
- Nesterov/FISTA provide optimal convergence rate.
- Generalization of inertial technique to first-order methods, or in general fixed-point iteration, is achievable:
  - Guaranteed sequence convergence.
  - **NO** acceleration guarantees. Unless stronger assumptions are imposed, *e.g.* strong convexity or Lipschitz smoothness.
- For a given method, *e.g.* Douglas–Rachford, the outcome its inertial/SOR versions is problem and **parameters** dependent.

A general acceleration framework with acceleration guarantees is missing!

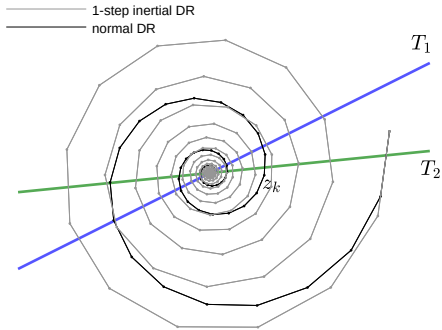
Trajectory of first-order methods

Straight line, logarithmic/elliptical spirals

What makes the difference?



Trajectory of $\{x_k\}_{k\in\mathbb{N}}$.

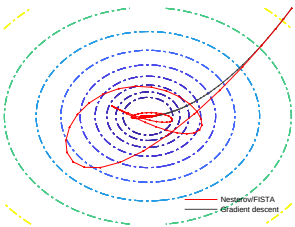


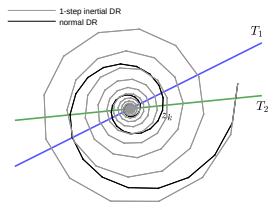
Trajectory of $\{z_k\}_{k\in\mathbb{N}}$.

`Projection of elliptical spiral in` $\mathbb{R}^4$ `onto` $\mathbb{R}^2$.

What makes the difference?

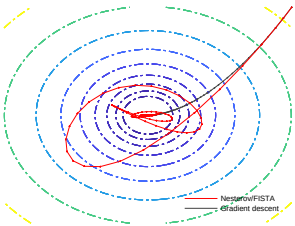
Non-smooth opt. ⟹ ☐ ⟹ ☐

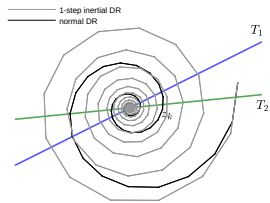




What makes the difference?

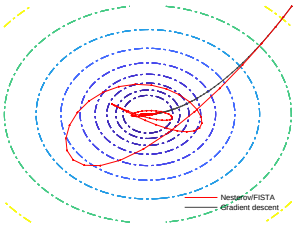
Non-smooth opt. ⟹ (Structure) **FoM** ⟹ [ ]

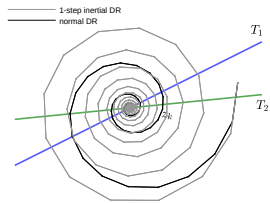




What makes the difference?

Non-smooth opt. $\xrightarrow{\text{Structure}}$ **FoM** $\xrightarrow{\text{Geometry}}$ **Trajectory of sequence**

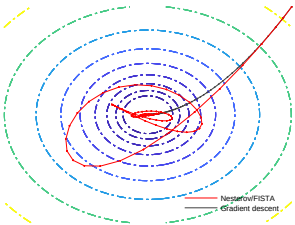


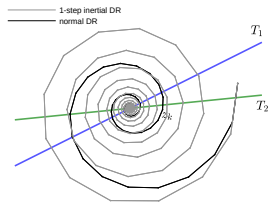


What makes the difference?

Non-smooth opt. ⟹ (Structure) **FoM** ⟹ (Geometry) **Trajectory of sequence**

However, FoM are **non-linear** in general...





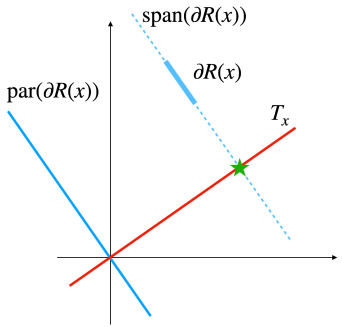
Partial smoothness

Definition - Partly smooth function [Lewis '03]

R is *partly smooth at x relative to a set $\mathcal{M}_x$* containing x if $\partial R(x) \neq \emptyset$ and

- Smoothness $\mathcal{M}_x$ is a C^2-manifold, $R|_{\mathcal{M}_x}$ is C^2 near x.
- Sharpness Tangent space $\mathcal{T}_{\mathcal{M}_x}(x)$ is $T_x \stackrel{\text{def}}{=} \mathrm{par}(\partial R(x))^{\perp}$.
- Continuity $\partial R : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is continuous along $\mathcal{M}_x$ near x.

$\mathrm{par}(C)$: sub-space parallel to C, where $C \subset \mathbb{R}^n$ is a non-empty convex set.



Examples:

- $\ell_1, \ell_{1,2}, \ell_\infty$-norm
- Nuclear norm
- Total variation
- Partly smooth sets...

Partial smoothness

Definition - Partly smooth function [Lewis '03]

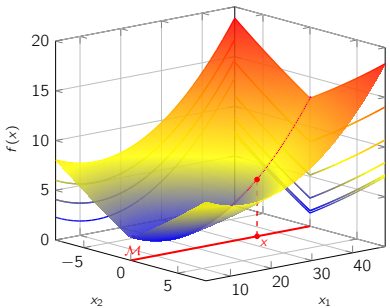
R is *partly smooth at* x *relative to a set* $\mathcal{M}_x$ containing x if $\partial R(x) \neq \emptyset$ and

Smoothness $\mathcal{M}_x$ is a C^2-manifold, $R|_{\mathcal{M}_x}$ is C^2 near x.

Sharpness Tangent space $\mathcal{T}_{\mathcal{M}_x}(x)$ is $T_x \stackrel{\text{def}}{=} \mathrm{par}(\partial R(x))^{\perp}$.

Continuity $\partial R : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is continuous along $\mathcal{M}_x$ near x.

$\mathrm{par}(C)$: sub-space parallel to C, where $C \subset \mathbb{R}^n$ is a non-empty convex set.



Examples:

- $\ell_1, \ell_{1,2}, \ell_\infty$-norm
- Nuclear norm
- Total variation
- Partly smooth sets...

Trajectory of first-order methods

Framework for analyzing the local trajectory of FoM

First-order method **(non-linear)**

$\Downarrow$

Convergence & Non-degeneracy: finite identification of $\mathcal{M}$

$\Downarrow$

Local linearization along $\mathcal{M}$: matrix M (linear)

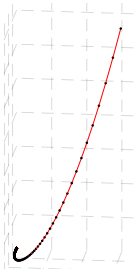
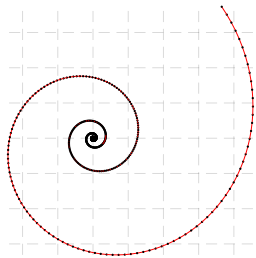
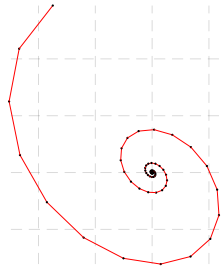
$\Downarrow$

Spectral properties of M

$\Downarrow$

Local trajectory

Trajectory of first-order methods

Forward–Backward	Douglas–Rachford/ADMM	Primal–Dual
M is similar to a **symmetric matrix** with real eigenvalues in $]-1,1]$.	Both functions are polyhedral, M is **normal** with complex eigenvalues of the form $\cos(\theta)\mathrm{e}^{\pm \mathrm{i}\theta}$.	Both functions are polyhedral, M is **block diagonalizable**.
		

NB: For DR/ADMM, if smoothness or (local) strong convexity is posed, **straight-line** trajectory can be obtained under proper parameters.

Adaptive acceleration for FoM

Trajectory following linear prediction

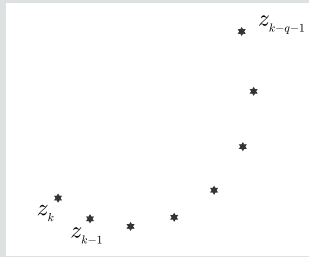
Linear prediction: illustration

Idea: Given past points $\{z_{k-j}\}_{j=0}^{q+1}$, how to predict z_{k+1}?

Define $\{v_{k-j} \stackrel{\text{def}}{=} z_{k-j} - z_{k-j-1}\}_{j=0}^{q}$.

- Fit v_k using past directions $v_{k-1}, \dots, v_{k-q}$:

$$c_k \stackrel{\text{def}}{=} \operatorname{argmin}_{c \in \mathbb{R}^q} \| \textstyle\sum_{j=1}^{q} c_j v_{k-j} - v_k \|^2.$$



Linear prediction: illustration

Idea: Given past points $\{z_{k-j}\}_{j=0}^{q+1}$, how to predict z_{k+1}?

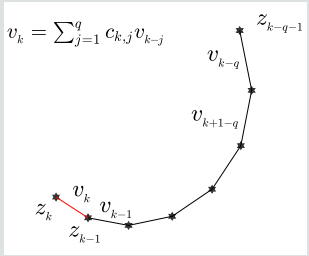
Define $\{v_{k-j} \overset{\text{def}}{=} z_{k-j} - z_{k-j-1}\}_{j=0}^{q}$.

- Fit v_k using past directions $v_{k-1}, \dots, v_{k-q}$:

$$c_k \overset{\text{def}}{=} \operatorname{argmin}_{c \in \mathbb{R}^q} \| \textstyle\sum_{j=1}^{q} c_j v_{k-j} - v_k \|^2.$$

- Suppose z_{k+1} is given, we can compute c_{k+1} as above...
- Approximation

$$v_{k+1} = \textstyle\sum_j c_{k+1,j} v_{k+1-j} \approx \sum_j c_{k,j} v_{k+1-j}.$$

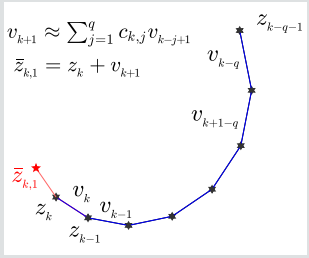


Linear prediction: illustration

Idea: Given past points $\{z_{k-j}\}_{j=0}^{q+1}$, how to predict z_{k+1}?

Define $\{v_{k-j} \overset{\text{def}}{=} z_{k-j} - z_{k-j-1}\}_{j=0}^{q}$.

- Approximation $z_{k+1} \approx \bar{z}_{k,1} \overset{\text{def}}{=} z_k + \sum_{j=1}^{q} c_{k,j} v_{k-j+1}$.

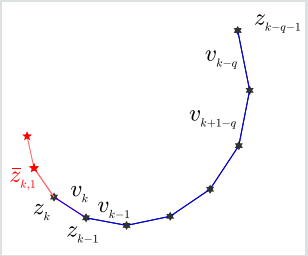


Linear prediction: illustration

Idea: Given past points $\{z_{k-j}\}_{j=0}^{q+1}$, how to predict z_{k+1}?

Define $\{v_{k-j} \overset{\text{def}}{=} z_{k-j} - z_{k-j-1}\}_{j=0}^{q}$.

- Approximation $z_{k+1} \approx \bar{z}_{k,1} \overset{\text{def}}{=} z_k + \sum_{j=1}^{q} c_{k,j} v_{k-j+1}$.
- Repeat on $\{z_{k-j}\}_{j=0}^{q} \cup \{\bar{z}_{k,1}\}$ and so on...

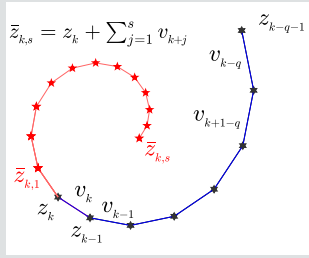


Linear prediction: illustration

Idea: Given past points $\{z_{k-j}\}_{j=0}^{q+1}$, how to predict z_{k+1}?

Define $\{v_{k-j} \overset{\text{def}}{=} z_{k-j} - z_{k-j-1}\}_{j=0}^{q}$.

- Approximation $z_{k+1} \approx \bar{z}_{k,1} \overset{\text{def}}{=} z_k + \sum_{j=1}^{q} c_{k,j} v_{k-j+1}$.
- Repeat on $\{z_{k-j}\}_{j=0}^{q} \cup \{\bar{z}_{k,1}\}$ and so on...
- Repeat s times: $z_{k+s} \approx \bar{z}_{k,s} = z_k + \sum_{j=1}^{s} v_{k+j}$.



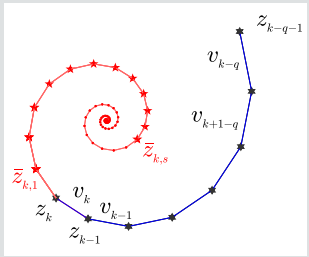
Linear prediction: illustration

Idea: Given past points $\{z_{k-j}\}_{j=0}^{q+1}$, how to predict z_{k+1}?

Define $\{v_{k-j} \overset{\text{def}}{=} z_{k-j} - z_{k-j-1}\}_{j=0}^{q}$.

- Approximation $z_{k+1} \approx \bar{z}_{k,1} \overset{\text{def}}{=} z_k + \sum_{j=1}^{q} c_{k,j} v_{k-j+1}$.
- Repeat on $\{z_{k-j}\}_{j=0}^{q} \cup \{\bar{z}_{k,1}\}$ and so on...
- Repeat s times: $z_{k+s} \approx \bar{z}_{k,s} = z_k + \sum_{j=1}^{s} v_{k+j}$.
- Let $s \to +\infty$, if converges

$$z^\star \approx \bar{z}_{k,+\infty} = z_k + \sum_{j=1}^{+\infty} v_{k+j}.$$



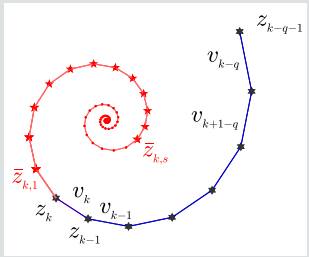
Linear prediction: illustration

Idea: Given past points $\{z_{k-j}\}_{j=0}^{q+1}$, how to predict z_{k+1}?

Define $\{v_{k-j} \overset{\text{def}}{=} z_{k-j} - z_{k-j-1}\}_{j=0}^{q}$.

- Approximation $z_{k+1} \approx \bar{z}_{k,1} \overset{\text{def}}{=} z_k + \sum_{j=1}^{q} c_{k,j} v_{k-j+1}$.
- Repeat on $\{z_{k-j}\}_{j=0}^{q} \cup \{\bar{z}_{k,1}\}$ and so on...
- Repeat s times: $z_{k+s} \approx \bar{z}_{k,s} = z_k + \sum_{j=1}^{s} v_{k+j}$.
- Let $s \to +\infty$, if converges

$$z^\star \approx \bar{z}_{k,+\infty} = z_k + \sum_{j=1}^{+\infty} v_{k+j}.$$



The ***s*-step extrapolation** is $\bar{z}_{k,s} = z_k + \mathcal{E}_{s,q,k}$, where $\mathcal{E}_{s,q,k} = \sum_{j=1}^{q} \hat{c}_j v_{k-j+1}$ and

$$\hat{c} \overset{\text{def}}{=} \left(\sum_{j=1}^{s} H(c^k)^j\right)_{(:,1)} \quad \text{with} \quad H(c^k) \overset{\text{def}}{=} \begin{bmatrix} c_1 & 1 & & \\ c_2 & & 1 & \\ \vdots & & & \ddots \\ c_{q-1} & & & 1 \\ c_q & 0 & \dots & 0 \end{bmatrix}.$$

Adaptive acceleration for FoM (A^{2}FoM)

Given first-order method

$$z_{k+1} = \mathcal{F}(z_k).$$

Algorithm - A^{2}FoM via linear prediction

Let $s \geq 1, q \geq 1$ be integers. Let $z_0 \in \mathbb{R}^n$ and $\bar{z}_0 = z_0$, set $D = 0 \in \mathbb{R}^{n\times(q+1)}$:

- For $k \geq 1$:

$$\begin{aligned} z_k &= \mathcal{F}(\bar{z}_{k-1}), \\ v_k &= z_k - z_{k-1}, \\ D &= [v_k, D(:, 1:q)]. \end{aligned}$$

- If $\mathrm{mod}(k, q+2) = 0$: compute c and H_c,

$$\textbf{If } \rho(H_c) < 1\textbf{: } \bar{z}_k = z_k + V_k(\textstyle\sum_{i=1}^{s} H_c^i)_{(:,1)}; \qquad \textbf{else: } \bar{z}_k = z_k.$$

  If $\mathrm{mod}(k, q+2) \neq 0$: $\bar{z}_k = z_k$.

Adaptive acceleration for FoM (A^{2}FoM)

Remarks

- Every $(q+2)$-iteration we apply 1-step LP.
- Only apply the linear prediction when $\rho(H_c) < 1$.
- Extra memory cost $n \times (q+1)$ (the difference vector matrix). Usually $q \leq 10$.
- Extra computation cost, $q^2 n$ from V_{k-1}^+.

Adaptive acceleration for FoM (A^{2}FoM)

Remarks

- Every $(q+2)$-iteration we apply 1-step LP.
- Only apply the linear prediction when $\rho(H_c) < 1$.
- Extra memory cost $n \times (q+1)$ (the difference vector matrix). Usually $q \leq 10$.
- Extra computation cost, $q^2 n$ from V_{k-1}^+.
- Conditional convergence can be obtained by treating LP as **perturbation error**,

$$z_{k+1} = \mathcal{F}(z_k + \epsilon_k).$$

- Weighted LP

$$\bar{z}_k = z_k + a_k V_k (\sum_{i=1}^{s} H_c^i)_{(:,1)},$$

  with a_k updated online.

Adaptive acceleration for FoM (A^{2}FoM)

Remarks

- Every $(q+2)$-iteration we apply 1-step LP.
- Only apply the linear prediction when $\rho(H_c) < 1$.
- Extra memory cost $n \times (q+1)$ (the difference vector matrix). Usually $q \leq 10$.
- Extra computation cost, $q^2 n$ from V_{k-1}^+.
- Conditional convergence can be obtained by treating LP as **perturbation error**,

$$z_{k+1} = \mathcal{F}(z_k + \epsilon_k).$$

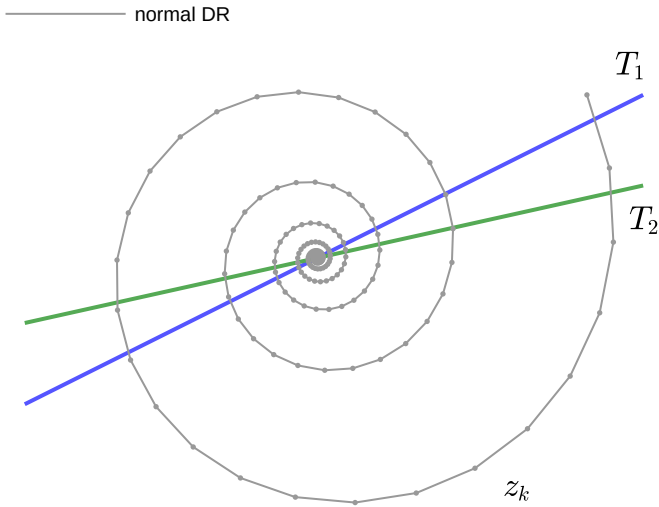
- Weighted LP

$$\bar{z}_k = z_k + a_k V_k (\sum\nolimits_{i=1}^{s} H_c^i)_{(:,1)},$$

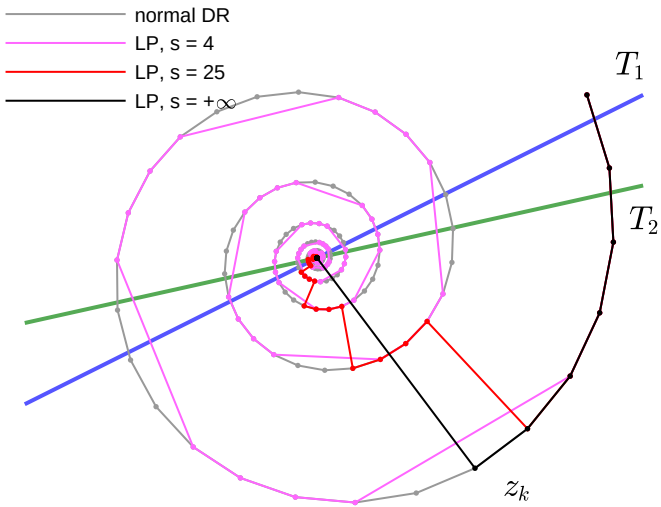
  with a_k updated online.
- If $\rho(H_c) < 1$, the Neumann series is convergent

$$\sum\nolimits_{i=0}^{+\infty} H_c^i = (\mathrm{Id} - H_c)^{-1}.$$

Example: Douglas–Rachford continue



Example: Douglas–Rachford continue



Relation with previous work

Polynomial extrapolation, regularized non-linear acceleration

Convergence acceleration

Given a sequence $\{z_k\}_{k\in\mathbb{N}}$ which converges to $z^\star$. Can we generate another sequence $\{\bar{z}_k\}_{k\in\mathbb{N}}$ such that $\|\bar{z}_k - z^\star\| = o(\|z_k - z^\star\|)$?

Convergence acceleration

Given a sequence $\{z_k\}_{k\in\mathbb{N}}$ which converges to $z^\star$. Can we generate another sequence $\{\bar{z}_k\}_{k\in\mathbb{N}}$ such that $\|\bar{z}_k - z^\star\| = o(\|z_k - z^\star\|)$?

This is called **convergence acceleration** and is well-established in numerical analysis:

- 1927 Aitkin's Δ-process.
- 1965 Andersen's acceleration.
- 1970's Vector extrapolation techniques such as minimal polynomial extrapolation (MPE) and reduced rank extrapolation (RRE) [Sidi '17].
- Recent Regularized non-linear acceleration (RNA) is a regularised version of RRE introduced by [Scieur, D'Aspremont, Bach '16].

Polynomial extrapolation [Cabay & Jackson '76]

Consider $z_{k+1} = Mz_k + d$ with $\rho(M) < 1$ such that $z_k \to z^\star$:

- $z_k - z^\star = M(z_{k-1} - z^\star) = M^k(z_0 - z^\star)$,
- If $P(\lambda) = \sum_{j=0}^{q} c_j \lambda^j$ is the minimal polynomial of M w.r.t. $z_0 - z^\star$, that is

$$P(M)(z_0 - z^\star) = \sum_{j=0}^{q} c_j M^j (z_0 - z^\star) = 0.$$

  then $z^\star = \frac{\sum_{j=0}^{q} c_j z_j}{\sum_j c_j}$.

- The coefficients c can be computed without knowledge of $z^\star$:

$$V_q c_{(0:q-1)} = -v_{q+1}$$

  where $V_q = [v_1|v_2|\cdots|v_q]$ and $v_j = z_j - z_{j-1}$.

Vector extrapolation techniques

Vector extrapolation methods with applications (SIAM, 2017) by Avram Sidi.

Minimal polynomial extrapolation (MPE)

Let $z_0 = \bar{z}$:

S.1 Generate points $\{z_j\}_{j=0}^{q+1}$ and let $v_j = z_j - z_{j-1}$.

S.2 Let $c \in \mathbb{R}^{q+1}$ be s.t. $c_q = 1$ and $V_q c_{(0:q-1)} = -v_{q+1}$ where $V_q = [v_1|\cdots|v_q]$. For $j \in [0, q-1]$, $\tilde{c}_j \overset{\text{def}}{=} c_j / (\sum_{j=0}^{q} c_j)$.

S.3 $\bar{z} \overset{\text{def}}{=} \sum_{j=0}^{q} \tilde{c}_j z_j$.

Vector extrapolation techniques

Vector extrapolation methods with applications (SIAM, 2017) by Avram Sidi.

Minimal polynomial extrapolation (MPE)

Let $z_0 = \bar{z}$:

S.1 Generate points $\{z_j\}_{j=0}^{q+1}$ and let $v_j = z_j - z_{j-1}$.

S.2 Let $c \in \mathbb{R}^{q+1}$ be s.t. $c_q = 1$ and $V_q c_{(0:q-1)} = -v_{q+1}$ where $V_q = [v_1|\cdots|v_q]$. For $j \in [0, q-1]$, $\tilde{c}_j \overset{\text{def}}{=} c_j / (\sum_{j=0}^{q} c_j)$.

S.3 $\bar{z} \overset{\text{def}}{=} \sum_{j=0}^{q} \tilde{c}_j z_j$.

Reduced rank extrapolation (RRE)

[Andersen '65; Kaniel & Stein '74; Eddy '79; Mešina '77] Replace step **S.2** by

$$\tilde{c} \in \operatorname{argmin}_c \|V_{q+1} c\| \quad \text{subject to} \quad \mathbf{1}^T c = 1.$$

Vector extrapolation techniques

Vector extrapolation methods with applications (SIAM, 2017) by Avram Sidi.

Minimal polynomial extrapolation (MPE)

Let $z_0 = \bar{z}$:

- S.1 Generate points $\{z_j\}_{j=0}^{q+1}$ and let $v_j = z_j - z_{j-1}$.
- S.2 Let $c \in \mathbb{R}^{q+1}$ be s.t. $c_q = 1$ and $V_q c_{(0:q-1)} = -v_{q+1}$ where $V_q = [v_1|\cdots|v_q]$. For $j \in [0, q-1]$, $\tilde{c}_j \overset{\text{def}}{=} c_j / (\sum_{j=0}^{q} c_j)$.
- S.3 $\bar{z} \overset{\text{def}}{=} \sum_{j=0}^{q} \tilde{c}_j z_j$.

Reduced rank extrapolation (RRE)

[Andersen '65; Kaniel & Stein '74; Eddy '79; Mešina '77] Replace step **S.2** by

$$\tilde{c} \in \operatorname{argmin}_c \|V_{q+1} c\| \quad \text{subject to} \quad \mathbf{1}^T c = 1.$$

LP is equivalent to MPE with **S.3** replaced by $\bar{z} \overset{\text{def}}{=} \sum_{j=0}^{q} \tilde{c}_j z_{j+1}$.

Relationship between MPE and LP

Our derivation

- is based on sequence **trajectory**.
- motivates checking $\rho(H_c) < 1$.

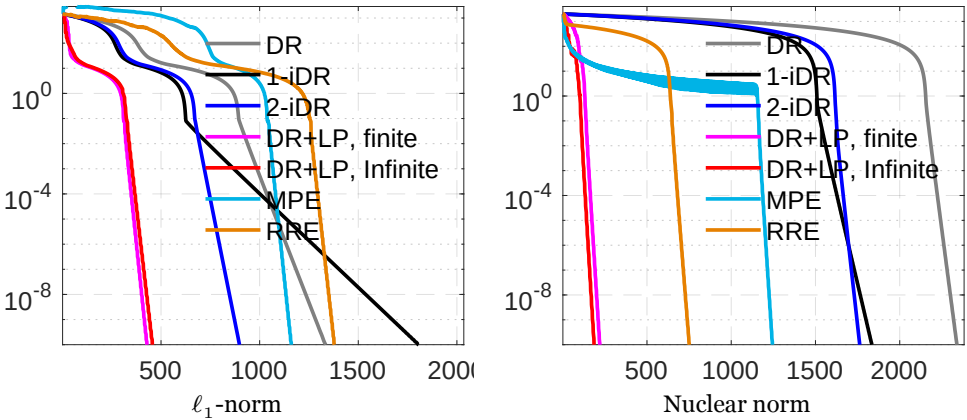
Relationship between MPE and LP

Our derivation

- is based on sequence **trajectory**.
- motivates checking $\rho(H_c) < 1$.

Example: Solve with DR

$$\min_x R(x) \text{ subject to } Ax = b.$$



Acceleration guarantees

When $z_{k+1} - z_k = M(z_k - z_{k-1})$,

$$\|\bar{z}_{k,s} - z^\star\| \leq \|z_{k+s} - z^\star\| + B\epsilon_k$$

where $\epsilon_k = \|V_{k-1}c - v_k\|$ and $B \stackrel{\text{def}}{=} \sum_{\ell=1}^{s} \|M^\ell\| |\sum_{i=0}^{s-\ell} (H_c^i)_{(1,1)}|$.

Acceleration guarantees

When $z_{k+1} - z_k = M(z_k - z_{k-1})$,

$$\|\bar{z}_{k,s} - z^\star\| \leq \|z_{k+s} - z^\star\| + B\epsilon_k$$

where $\epsilon_k = \|V_{k-1}c - v_k\|$ and $B \stackrel{\text{def}}{=} \sum_{\ell=1}^{s} \|M^\ell\| |\sum_{i=0}^{s-\ell} (H_c^i)_{(1,1)}|$.

Asymptotic bound ($k \to \infty$):

$$\epsilon_k = O(|\lambda_{q+1}|^k)$$

where λ_{q+1} is the $(q+1)^{th}$ largest eigenvalue. Without extrapolation, we just have $O(|\lambda_1|^k)$.

Acceleration guarantees

When $z_{k+1} - z_k = M(z_k - z_{k-1})$,

$$\|\bar{z}_{k,s} - z^\star\| \leq \|z_{k+s} - z^\star\| + B\epsilon_k$$

where $\epsilon_k = \|V_{k-1}c - v_k\|$ and $B \stackrel{\text{def}}{=} \sum_{\ell=1}^{s} \|M^\ell\| \|\sum_{i=0}^{s-\ell} (H_c^i)_{(1,1)}|$.

Asymptotic bound ($k \to \infty$):

$$\epsilon_k = O(|\lambda_{q+1}|^k)$$

where λ_{q+1} is the $(q+1)^{th}$ largest eigenvalue. Without extrapolation, we just have $O(|\lambda_1|^k)$.

Non-asymptotic bound: If $\Sigma(M) \subset [\alpha, \beta]$ with $-1 < \alpha < \beta < 1$, then

$$B\epsilon_k \leq K\beta^{k-q} \left(\frac{\sqrt{\eta} - 1}{\sqrt{\eta} + 1} \right)^q, \quad \text{where} \quad \eta = \frac{1-\alpha}{1-\beta}$$

Acceleration guarantees

When $z_{k+1} - z_k = M(z_k - z_{k-1})$,

$$\|\bar{z}_{k,s} - z^\star\| \leq \|z_{k+s} - z^\star\| + B\epsilon_k$$

where $\epsilon_k = \|V_{k-1}c - v_k\|$ and $B \stackrel{\text{def}}{=} \sum_{\ell=1}^{s} \|M^\ell\| \|\sum_{i=0}^{s-\ell} (H_c^i)_{(1,1)}|$.

Asymptotic bound ($k \to \infty$):

$$\epsilon_k = O(|\lambda_{q+1}|^k)$$

where λ_{q+1} is the $(q+1)^{th}$ largest eigenvalue. Without extrapolation, we just have $O(|\lambda_1|^k)$.

Non-asymptotic bound: If $\Sigma(M) \subset [\alpha, \beta]$ with $-1 < \alpha < \beta < 1$, then

$$B\epsilon_k \leq K\beta^{k-q} \left(\frac{\sqrt{\eta}-1}{\sqrt{\eta}+1} \right)^q, \quad \text{where} \quad \eta = \frac{1-\alpha}{1-\beta}$$

- Similar error bounds also hold for MPE and RRE [Sidi '98].

Acceleration guarantees

When $z_{k+1} - z_k = M(z_k - z_{k-1})$,

$$\|\bar{z}_{k,s} - z^\star\| \leq \|z_{k+s} - z^\star\| + B\epsilon_k$$

where $\epsilon_k = \|V_{k-1}c - v_k\|$ and $B \stackrel{\text{def}}{=} \sum_{\ell=1}^{s} \|M^\ell\| \left|\sum_{i=0}^{s-\ell} (H_c^i)_{(1,1)}\right|$.

Asymptotic bound ($k \to \infty$):

$$\epsilon_k = O(|\lambda_{q+1}|^k)$$

where λ_{q+1} is the $(q+1)^{th}$ largest eigenvalue. Without extrapolation, we just have $O(|\lambda_1|^k)$.

Non-asymptotic bound: If $\Sigma(M) \subset [\alpha, \beta]$ with $-1 < \alpha < \beta < 1$, then

$$B\epsilon_k \leq K\beta^{k-q} \left(\frac{\sqrt{\eta}-1}{\sqrt{\eta}+1}\right)^q, \quad \text{where} \quad \eta = \frac{1-\alpha}{1-\beta}$$

- Similar error bounds also hold for MPE and RRE [Sidi '98].
- For PD, DR with polyhedral functions, guaranteed acceleration with $q = 2$.

Regularised non-linear acceleration (RNA): regularised RRE

[Scieur, D'Aspremont, Bach '16] studied RRE for the case of

$$z_{k+1} - z^\star = A(z_k - z^\star) + O(\|z_k - z^\star\|^2)$$

where A is **symmetric** with $0 \preceq A \preceq \sigma \mathrm{Id}$, $\sigma < 1$.

Regularised non-linear acceleration (RNA): regularised RRE

[Scieur, D'Aspremont, Bach '16] studied RRE for the case of

$$z_{k+1} - z^\star = A(z_k - z^\star) + O(\|z_k - z^\star\|^2)$$

where A is **symmetric** with $0 \preceq A \preceq \sigma \mathrm{Id}$, $\sigma < 1$.

To deal with the possible ill-conditioning of V_q, regularise with $\lambda > 0$:

$$\tilde{c} \in \mathrm{Argmin}_c \ \|c^T(V_q^T V_q + \lambda \mathrm{Id})c\| \quad \text{subject to} \quad \mathbf{1}^T c = 1.$$

Regularised non-linear acceleration (RNA): regularised RRE

[Scieur, D'Aspremont, Bach '16] studied RRE for the case of

$$z_{k+1} - z^\star = A(z_k - z^\star) + O(\|z_k - z^\star\|^2)$$

where A is **symmetric** with $0 \preceq A \preceq \sigma \mathrm{Id}$, $\sigma < 1$.

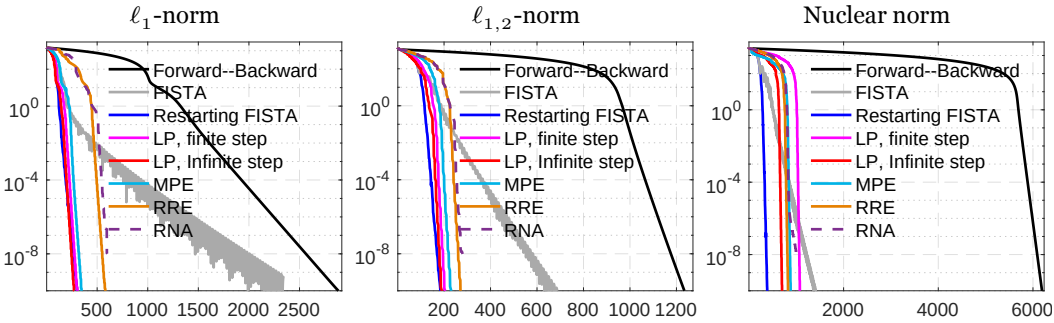
To deal with the possible ill-conditioning of V_q, regularise with $\lambda > 0$:

$$\tilde{c} \in \mathrm{Argmin}_c \ \|c^T(V_q^T V_q + \lambda \mathrm{Id})c\| \quad \text{subject to} \quad \mathbf{1}^T c = 1.$$

- In practice, grid search on optimal $\lambda \in [\lambda_{\min}, \lambda_{\max}]$. Potentially many evaluations of the objective.
- The angle between $z_k - z_{k-1}$ and $z_{k+1} - z_k$ converges to zero, intuitively, this is the regime where standard inertial works well...

LASSO-type problems with FB

NB: Recall the straight-line trajectory of FB.



Note the performance of restarted-FISTA [O'Donoghue & Candès '12]!

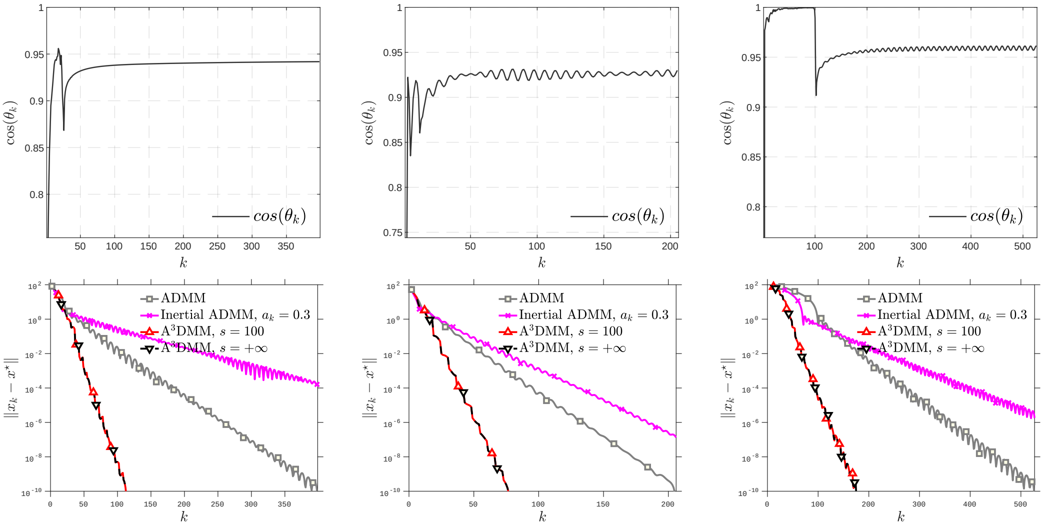
Numerical experiments

Some more results

Experiment: 2 non-smooth terms

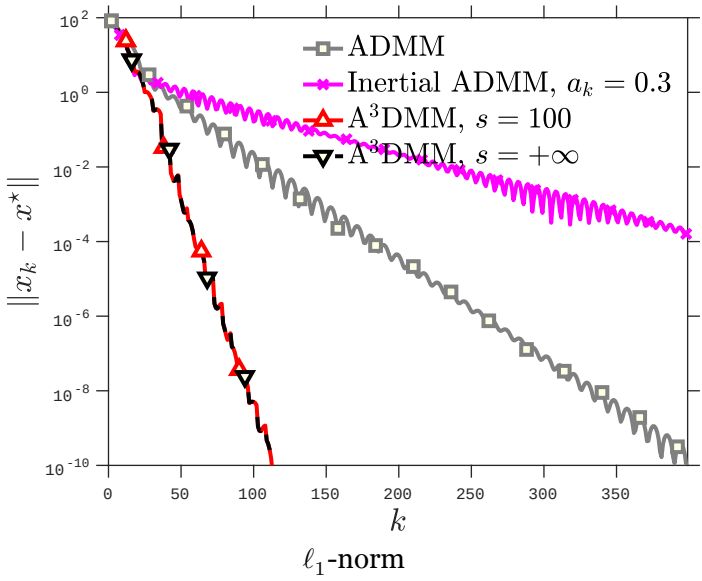
Basis pursuit type problem with $\Omega \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : Kx = f\}$:

$$\min_{x,y\in\mathbb{R}^n} \; R(x) + \iota_\Omega(y) \quad \text{such that} \quad x - y = 0.$$



ℓ_1-norm $\qquad$ $\ell_{1,2}$-norm $\qquad$ Nuclear norm

Experiment: 2 non-smooth terms



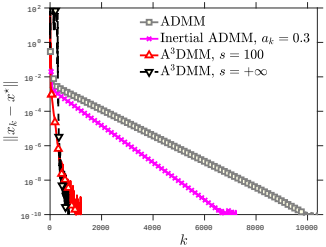
ℓ_1-norm

Inertial ADMM is **slower** than ADMM as eventual trajectory is a spiral.

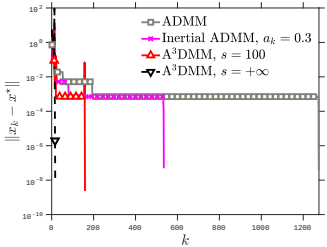
Experiment: LASSO

The LASSO problem

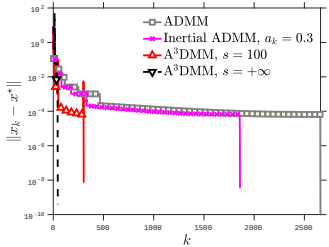
$$\min_{x,y\in\mathbb{R}^n} R(x) + \frac{1}{2}\|Ky - f\|^2 \quad \text{such that} \quad x - y = 0.$$



covtype

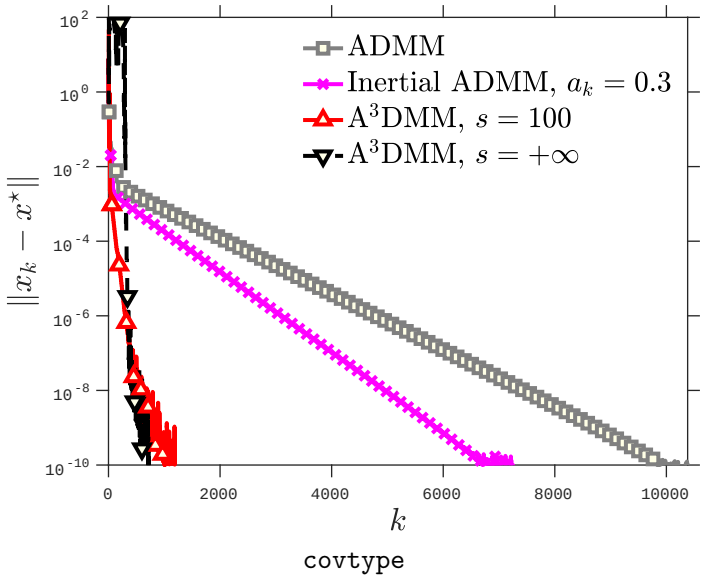


ijcnn1



phishing

Experiment: LASSO



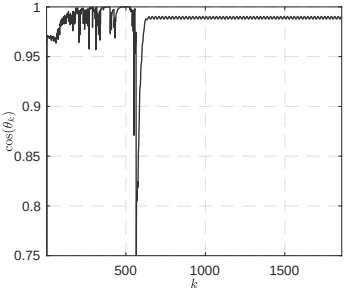
Inertial ADMM does accelerate, but A^{3}DMM is significantly faster.

Experiment: Total variation based image inpainting

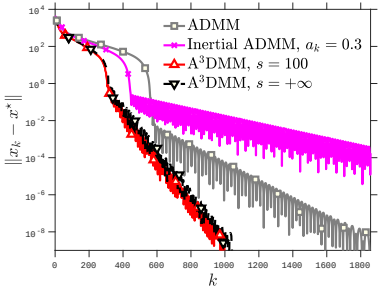
Let $\Omega \overset{\text{def}}{=} \{x \in \mathbb{R}^{n\times n} : P_{\mathcal{D}}(x) = f\}$, $P_{\mathcal{D}}$ randomly sets 50% pixels to zero and consider

$$\min_{x\in\mathbb{R}^{n\times n}} \|y\|_1 + \iota_{\Omega}(x) \quad \text{such that} \quad \nabla x - y = 0.$$

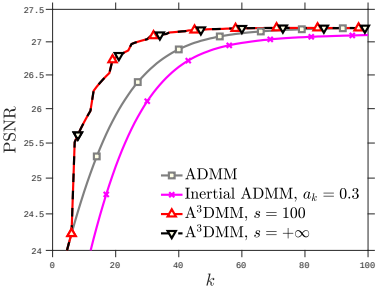
Angle $\{\theta_k\}$



$\|x_k - x^\star\|$



PSNR



- Both functions are polyhedral, trajectory is a spiral.
- Inertial ADMM is **slower** than ADMM.

Experiment: Total variation based image inpainting

Original image

ADMM, PSNR = 26.5448

Inertial ADMM, PSNR = 26.1096

Corrupted image

A^{3}DMM $s = 100$, PSNR = 27.0402

A^{3}DMM $s = +\infty$, PSNR = 27.0402

Take-away messages

Trajectory of FoM For fixed-point sequence $\{z_k\}_{k\in\mathbb{N}}$

- Linearization of FoM.
- Locally different FoM demonstrate **distinct trajectories**: straight line, (logarithmic and elliptic) spirals.

Take-away messages

Trajectory of FoM For fixed-point sequence $\{z_k\}_{k\in\mathbb{N}}$

- Linearization of FoM.
- Locally different FoM demonstrate **distinct trajectories**: straight line, (logarithmic and elliptic) spirals.

An adaptive acceleration for FoM

- Though motivated by local trajectory, linear prediction works **globally**.
- For polyhedral functions, guaranteed local acceleration for DR/PD using 4 past points.
- For FB, the trajectory is eventually a straight line and one can guarantee local acceleration by extrapolating from 3 past points under our framework, but one could just apply **restarted FISTA**...

Take-away messages

Trajectory of FoM For fixed-point sequence $\{z_k\}_{k\in\mathbb{N}}$

- Linearization of FoM.
- Locally different FoM demonstrate **distinct trajectories**: straight line, (logarithmic and elliptic) spirals.

An adaptive acceleration for FoM

- Though motivated by local trajectory, linear prediction works **globally**.
- For polyhedral functions, guaranteed local acceleration for DR/PD using 4 past points.
- For FB, the trajectory is eventually a straight line and one can guarantee local acceleration by extrapolating from 3 past points under our framework, but one could just apply **restarted FISTA**...

Reference

- `Trajectory of Alternating Direction Method of Multipliers and Adaptive Acceleration, NeurIPS19.`
- `Geometry of First-Order Methods and Adaptive Acceleration, arXiv:2003.03910.`

Many thanks for your time!

https://jliang993.github.io/