Monotonicity, Acceleration, Inertia, and the Proximal Gradient algorithm

Franck Iutzeler LJK, Université Grenoble Alpes

OSL, Les Houches April 10, 2017

Problem: Solving $\min_x F(x)$

Method $x_{k+1} = \mathcal{M}(x_k)$ (deterministic non-linear operation)

Operator viewpoint:

contraction properties
$\|\mathcal{M}(x) - \mathcal{M}(y)\| \leq \|x - y\|$

of the *iterates* (x_k)

towards *fixed points* $x^\star$

Optimization viewpoint:

descent properties
$F(\mathcal{M}(x)) - F(x) \leq -\|\mathcal{M}(x) - x\|$

of the *functional values* $(F(x_k))$

towards *minimizers* $F^\star$

Algorithm Acceleration: speeding up our method of choice $\mathcal{M}$ for a *small computational cost* compared to $\mathcal{M}$

- ~~Newton's method~~ $x_{k+1} = \mathcal{N} \circ \mathcal{M}(x_k)$
- Damping/*Relaxation* $x_{k+1} = \mathcal{M}(x_k) + (\eta - 1)(\mathcal{M}(x_k) - x_k)$
- Nesterov/Fast/*Inertia* $x_{k+1} = \mathcal{M}(x_k) + \gamma(\mathcal{M}(x_k) - \mathcal{M}(x_{k-1}))$

- Acceleration & Operators
- in Practice
- Bridging Relaxation & Inertia
- the Proximal Gradient algorithm

Acceleration & Operators

in Practice

Bridging Relaxation & Inertia

the Proximal Gradient algorithm

Firm non-expansivity: *The fixed point method $\mathcal{M}$ is firmly non-expansive if for any fixed point $x^\star$ and any x*

$$\|\mathcal{M}(x) - x^\star\|^2 \leq \|x - x^\star\|^2 - \|\mathcal{M}(x) - x\|^2.$$

Convergence theorem [Krasnoselskiĭ,1955-Mann,1953]
Let $\mathcal{M}$ be firmly non-expansive with fixed points, then the iterations

$$x_{k+1} = \mathcal{M}(x_k)$$

converge to a fixed point of $\mathcal{M}$.

- *Fejér* monotonous $\|x_{k+1} - x^\star\|^2 \leq \|x_k - x^\star\|^2$
- $O(1/k)$ in general
- Linear under additional assumptions (strong convexity, polyhedral)
- Encompasses
  - From a simple gradient with $\gamma \leq 1/L$ stepsize [Baillon-Haddad,1977]
  - to ADMM [Lions-Mercier,1979]
  - and more complex methods [Chambolle-Pock,2011;Condat,2013;...]

$$\begin{cases} y_{k+1} = \mathcal{M}(x_k) \\ \mathbf{x_{k+1}} = \cancel{y_{k+1}} \ \textbf{extrapolation}\left(\mathbf{y_{k+1}}, (\mathbf{y_k}), (\mathbf{x_k})\right) \end{cases}$$

Assumption:

The fixed point method $\mathcal{M}$ is firmly non-expansive i.e. for any fixed point $x^\star$ and any x

$$\|\mathcal{M}(x) - x^\star\|^2 \leq \|x - x^\star\|^2 - \|\mathcal{M}(x) - x\|^2.$$

Acceleration:

Using
- operation output y_{k+1}
- past outputs $y_k, y_{k-1}, \ldots$
- past iterates $x_k, x_{k-1}, \ldots$

to find a *better* point x_{k+1} than y_{k+1}

Two main strategies:

- Relaxation $\mathbf{x_{k+1}} = \mathbf{y_{k+1}} + (\eta_\mathbf{k} - \mathbf{1})(\mathbf{y_{k+1}} - \mathbf{x_k})$

  plays on the methods contraction.
- Inertia $\mathbf{x_{k+1}} = \mathbf{y_{k+1}} + \gamma_\mathbf{k}(\mathbf{y_{k+1}} - \mathbf{y_k})$

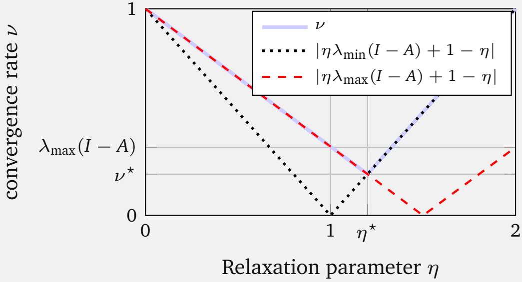
  plays on the moments of the iterates sequence.

Richardson iterations (1910): Solve linear systems by linear updates

$$x^{k+1} = x^k - (Ax^k - b) + \eta(Ax^k - b)$$

- Faster linear (exponential) convergence rate for chosen η
- Optimal η gives Chebyshev iterations

$$\eta = 1 + \frac{2}{\lambda_{\min}(A) + \lambda_{\max}(A)}$$



Krasnoselskiĭ–Mann iterations (1955): Relaxation is present in the operator convergence theorem.

$$\begin{cases} y_{k+1} = \mathcal{M}(x_k) \\ \mathbf{x_{k+1} = y_{k+1} + (\eta_{k+1} - 1)(y_{k+1} - x_k)} \end{cases}$$ with $\mathcal{M}$ firmly non-expansive

Relaxation *converges if* $0 < \liminf \eta_k \leq \limsup \eta_k < 2$.

- Fejér monotonous $\|x_{k+1} - x^\star\| \leq \|x_k - x^\star\|$
- Limit case: $\mathcal{M}([x, y]) = [x, 0]$. Take $\eta = 2$, then $\mathcal{M}_\eta([x, y]) = [x + 0, 0 + (-y)] = [x, -y]$

---

gradient algorithm:

$$x^{k+1} = x^k - \frac{\eta_{k+1}}{L} \nabla f(x_k)$$

- *"optimal"* $\frac{2}{1+\mu/L}$ with μ-strong convexity

ADMM:

Update is more involved (see later)

- *"$\eta \in [1.5, 1.8]$ usually speeds up the convergence"* [Eckstein'92]

- [Giselsson-Falk-Boyd'16] proposed a line search to compute an η_k that sufficiently decrease the residual

Acceleration & Operators

Relaxation **Inertia**

in Practice

Relaxation Inertia Application

Bridging Relaxation & Inertia

Intuition Alt. Inertia

the Proximal Gradient algorithm

Acceleration Alt. inertia

Fast gradient of Nesterov (1983): *optimal* first order method for minimizing an L-smooth convex function f

$$\begin{cases} y_{k+1} = x_k - \frac{1}{L}\nabla f(x_k) \\ x_{k+1} = y_{k+1} + \gamma_{k+1}(y_{k+1} - y_k) \end{cases}$$

with $\gamma_{k+1} = \frac{t_k - 1}{t_{k+1}} \to 1$ where $t_0 = 0$ and $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$.

FISTA (2008): fast proximal gradient method for minimizing an L-smooth convex function f plus a convex function g

$$\begin{cases} y_{k+1} = \arg\min_x \left\{ g(x) + \frac{L}{2}\left\| x - (x_k - \frac{1}{L}\nabla f(x_k)) \right\|^2 \right\} \\ x_{k+1} = y_{k+1} + \gamma_{k+1}(y_{k+1} - y_k) \end{cases}$$

- Faster (sub-linear) convergence rate: $\mathcal{O}(1/k) \to \mathcal{O}(1/k^2)$

Differential inclusion viewpoint: $\dot{x}(t) = -\nabla f(x(t))$

- Explicit/Euler scheme: $\frac{x_{k+1}-x_k}{h} = -\nabla f(x^k) \Rightarrow x_{k+1} = x_k - h\nabla f(x^k)$

adding a second order term: $\ddot{x}(t) + \alpha(t)\dot{x}(t) = -\nabla f(x(t))$

$$\frac{x_{k+2} - 2x_{k+1} + x_k}{h^2} + \alpha_k \frac{x_{k+1} - x_k}{h} = -\nabla f(y_{k+1})$$

$$x_{k+2} = \underbrace{x_{k+1} + (1 - h\alpha_k)(x_{k+1} - x_k)}_{y_{k+1}} - h^2 \nabla f(y_{k+1})$$

- $\alpha(t) = \alpha \rightarrow$ fixed inertia; $\alpha(t) = \alpha/t \rightarrow \quad \gamma_k = \frac{k-1}{k+\alpha-1}$.
- Used recently [Attouch'15] to prove iterates convergence of accelerated Forward-Backward

---

Geometric viewpoint: see S. Bubeck's blog and [Bubeck et al.'15]

Last week: "*Why momentum really works*" by G. Goh at `http://distill.pub/2017/momentum/`

$$\begin{cases} y_{k+1} = \mathcal{M}(x_k) \\ \mathbf{x_{k+1} = y_{k+1} + \gamma_{k+1}(y_{k+1} - y_k)} \end{cases}$$ with $\mathcal{M}$ firmly non-expansive

Inertia *converges if* $\limsup \gamma_k < 1/3$

- Not Fejér monotonous
- Limit case: $T = 0.5I + 0.5 \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$.

---

gradient algorithm:

$$\begin{cases} y_{k+1} = x_k - \frac{1}{L}\nabla f(x_k) \\ x_{k+1} = y_{k+1} + \gamma_{k+1}(y_{k+1} - y_k) \end{cases}$$

- *"optimal"* $\frac{1-\sqrt{\mu/L}}{1+\sqrt{\mu/L}}$ with μ-strong convexity

ADMM:

Update is more involved (see later)

- ADMM + Nesterov sequence *on top* = Fast ADMM [Golstein et al.'14] but cv. by restart

---

- [Lin-Harchaoui-Mairal,'15+'17] Inertia-based *double-loop* Catalyst for opt.
- [Flammarion-Bach,'15] Links between averaging and inertia

Acceleration & Operators

- in Practice

Bridging Relaxation & Inertia

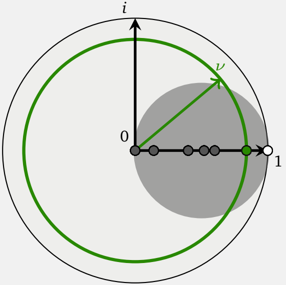
the Proximal Gradient algorithm

Goal: building a *simple* acceleration method from

- *contraction* property verified by the method
  Firmly non-expansive $\|\mathcal{M}(x) - x^\star\|^2 \leq \|x - x^\star\|^2 - \|\mathcal{M}(x) - x\|^2$
- *relaxation* or *inertia*
  as seen before
- accelerate the *linear rate*
  without knowledge of *strong*-*
  better adaptation to local properties and easily attained in practice

Affine approximation: $\mathcal{M}(x) = Rx + d$
where R is a symmetric matrix and d a vector of matching size.

- *contraction*
  $\Rightarrow$ eigs. are in the grey disk
- *linear rate* ν
  $\|x_k - x^\star\| = \tilde{O}(\nu^k)$



eigenvalues of R

- Effect of *relaxation/inertia*

Acceleration & Operators

Relaxation Inertia

in Practice

Relaxation Inertia Application

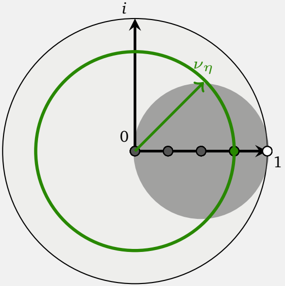
Bridging Relaxation & Inertia

Intuition Alt. Inertia

the Proximal Gradient algorithm

Acceleration Alt. inertia

$$\begin{cases} y_{k+1} = Rx_k + d \\ x_{k+1} = y_{k+1} + (\eta - 1)(y_{k+1} - x_k) \end{cases} \Rightarrow \mathbf{R}_\eta = \eta\mathbf{R} + (\mathbf{1} - \eta)\mathbf{I} \quad \text{on } x_k$$



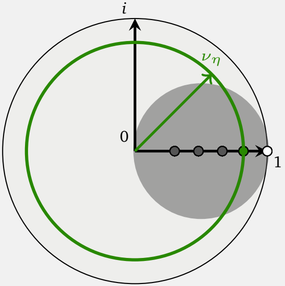
eigenvalues of R_η

- $\eta = 1$
- $\nu_\eta = 0.75$

$$\eta^\star = \frac{2}{2-\nu}$$
$$\nu^\star = \frac{\nu}{2-\nu}$$

- Depends on *extremal* eigenvalues
- Worst case at rate ν : $[0, \nu]$
- In this example $\nu = 0.75$

$$\begin{cases} y_{k+1} = Rx_k + d \\ x_{k+1} = y_{k+1} + (\eta - 1)(y_{k+1} - x_k) \end{cases} \Rightarrow \mathbf{R}_\eta = \eta\mathbf{R} + (\mathbf{1} - \eta)\mathbf{I} \quad \text{on } x_k$$



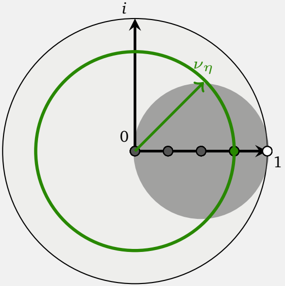
eigenvalues of R_η

- $\eta = 0.7$
- $\nu_\eta = 0.82$

$$\eta^\star = \frac{2}{2-\nu}$$
$$\nu^\star = \frac{\nu}{2-\nu}$$

- Depends on *extremal* eigenvalues
- Worst case at rate ν : $[0, \nu]$
- In this example $\nu = 0.75$

$$\begin{cases} y_{k+1} = Rx_k + d \\ x_{k+1} = y_{k+1} + (\eta - 1)(y_{k+1} - x_k) \end{cases} \Rightarrow \mathbf{R}_\eta = \eta\mathbf{R} + (\mathbf{1} - \eta)\mathbf{I} \quad \text{on } x_k$$



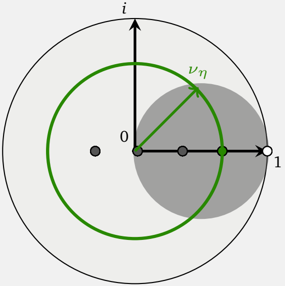
eigenvalues of R_η

- $\eta = 1$
- $\nu_\eta = 0.75$

$$\eta^\star = \frac{2}{2-\nu}$$
$$\nu^\star = \frac{\nu}{2-\nu}$$

- Depends on *extremal* eigenvalues
- Worst case at rate ν : $[0, \nu]$
- In this example $\nu = 0.75$

$$\begin{cases} y_{k+1} = R x_k + d \\ x_{k+1} = y_{k+1} + (\eta - 1)(y_{k+1} - x_k) \end{cases} \Rightarrow \mathbf{R}_\eta = \eta \mathbf{R} + (\mathbf{1} - \eta)\mathbf{I} \quad \text{on } x_k$$



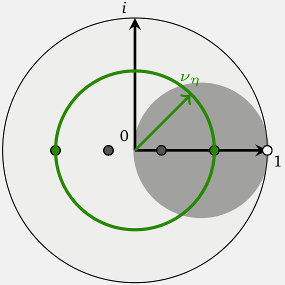
eigenvalues of R_η

- $\eta = 1.3$
- $\nu_\eta = 0.675$

$$\eta^\star = \frac{2}{2-\nu}$$
$$\nu^\star = \frac{\nu}{2-\nu}$$

- Depends on *extremal* eigenvalues
- Worst case at rate ν : $[0, \nu]$
- In this example $\nu = 0.75$

$$\begin{cases} y_{k+1} = Rx_k + d \\ x_{k+1} = y_{k+1} + (\eta - 1)(y_{k+1} - x_k) \end{cases} \Rightarrow \mathbf{R}_\eta = \eta\mathbf{R} + (\mathbf{1} - \eta)\mathbf{I} \quad \text{on } x_k$$



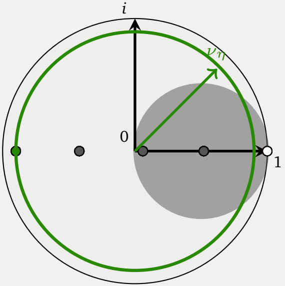
eigenvalues of R_η

- $\eta = 1.6 = \eta^\star$
- $\nu_\eta = 0.6 = \nu^\star$

$$\eta^\star = \frac{2}{2-\nu}$$
$$\nu^\star = \frac{\nu}{2-\nu}$$

- Depends on *extremal* eigenvalues
- Worst case at rate ν : $[0, \nu]$
- In this example $\nu = 0.75$

$$\begin{cases} y_{k+1} = Rx_k + d \\ x_{k+1} = y_{k+1} + (\eta - 1)(y_{k+1} - x_k) \end{cases} \Rightarrow \mathbf{R}_\eta = \eta\mathbf{R} + (\mathbf{1} - \eta)\mathbf{I} \quad \text{on } x_k$$



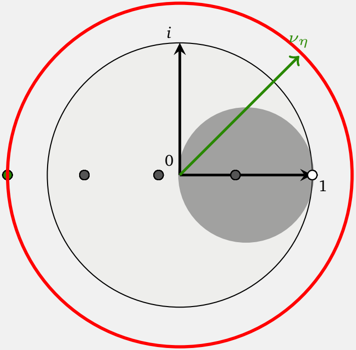
eigenvalues of R_η

- $\eta = 1.9$
- $\nu_\eta = 0.9$

$$\eta^\star = \frac{2}{2-\nu}$$
$$\nu^\star = \frac{\nu}{2-\nu}$$

- Depends on *extremal* eigenvalues
- Worst case at rate $\nu : [0, \nu]$
- In this example $\nu = 0.75$

$$\begin{cases} y_{k+1} = Rx_k + d \\ x_{k+1} = y_{k+1} + (\eta - 1)(y_{k+1} - x_k) \end{cases} \Rightarrow \mathbf{R}_\eta = \eta\mathbf{R} + (\mathbf{1} - \eta)\mathbf{I} \quad \text{on } x_k$$



- $\eta = 2.3$
- $\nu_\eta = 1.3$

eigenvalues of R_η

$$\eta^\star = \frac{2}{2-\nu}$$
$$\nu^\star = \frac{\nu}{2-\nu}$$

- Depends on *extremal* eigenvalues
- Worst case at rate $\nu : [0, \nu]$
- In this example $\nu = 0.75$

At an iteration $k > 2$,

- we know $x_k, x_{k-1}, \ldots, \eta_k, \eta_{k-1}, \ldots$

1. Estimate current rate $v_k = \frac{\eta_{k-1}\|x_k - x_{k-1}\|}{\eta_k\|x_{k-1} - x_{k-2}\|}$
2. Virtual eigenvalue $v_k = \eta_k \boldsymbol{\nu}_k + (1 - \eta_k) \Rightarrow \boldsymbol{\nu}_k = \frac{v_k - 1 + \eta_k}{\eta_k}$
3. Optimal relaxation on $\boldsymbol{\nu}_k$, $\eta_{k+1} = \frac{2}{2 - \boldsymbol{\nu}_k} = \frac{2\eta_k}{\eta_k + 1 - v_k}$

Online Relaxation for a FNE operator $\mathcal{M}$:

$$\eta_{k+1} = \frac{(2-\varepsilon)\eta_k}{\eta_k + 1 - \frac{\eta_{k-1}\|x_k - x_{k-1}\|}{\eta_k\|x_{k-1} - x_{k-2}\|}} + \frac{\varepsilon}{2}$$

$$x_{k+1} = \mathcal{M}(x_k) + (\eta_{k+1} - 1)(\mathcal{M}(x_k) - x_k)$$

- v_k is simplistic but theoretically consistent rate approx. as $v_k \in [0, 1]$
- we prove that $\eta_k \in [\frac{\varepsilon}{2}; 2 - \frac{\varepsilon}{2}]$ ensuring convergence for any FNE operator
- model inaccuracy is compensated by a constant re-estimation

Acceleration & Operators

Relaxation Inertia

in Practice

Relaxation **Inertia** Application

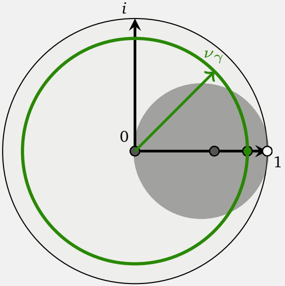
Bridging Relaxation & Inertia

Intuition Alt. Inertia

the Proximal Gradient algorithm

Acceleration Alt. inertia

$$\begin{cases} y_{k+1} = Rx^k + d \\ x_{k+1} = y_{k+1} + \gamma(y_{k+1} - y_k) \end{cases} \Rightarrow \mathbf{R}^\gamma = \begin{bmatrix} (1+\gamma)R & -\gamma R \\ I & 0 \end{bmatrix} \text{ on } \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix}$$



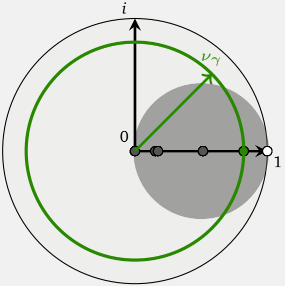
eigenvalues of R^γ

- $\gamma = 0$
- $\nu_\gamma = 0.85$

$$\gamma^\star = \frac{(1-\sqrt{1-\nu})^2}{\nu}$$
$$\nu^\star = 1 - \sqrt{1-\nu}$$

- Depends on the *largest* eigenvalue
- Worst case at rate ν : ν
- In this example $\nu = 0.85$

$$\begin{cases} y_{k+1} = Rx^k + d \\ x_{k+1} = y_{k+1} + \gamma(y_{k+1} - y_k) \end{cases} \Rightarrow \mathbf{R}^\gamma = \begin{bmatrix} (1+\gamma)R & -\gamma R \\ I & 0 \end{bmatrix} \text{ on } \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix}$$



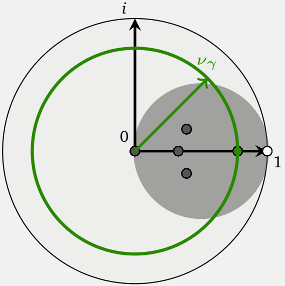
eigenvalues of R^γ

- $\gamma = 0.15$
- $\nu_\gamma = 0.822$

$$\gamma^\star = \frac{(1-\sqrt{1-\nu})^2}{\nu}$$
$$\nu^\star = 1 - \sqrt{1-\nu}$$

- Depends on the *largest* eigenvalue
- Worst case at rate $\nu : \nu$
- In this example $\nu = 0.85$

$$\begin{cases} y_{k+1} = Rx^k + d \\ x_{k+1} = y_{k+1} + \gamma(y_{k+1} - y_k) \end{cases} \Rightarrow \mathbf{R}^\gamma = \begin{bmatrix} (1+\gamma)R & -\gamma R \\ I & 0 \end{bmatrix} \text{ on } \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix}$$



eigenvalues of R^γ

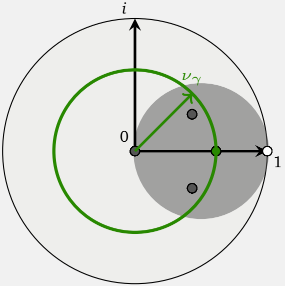
- $\gamma = 0.3$
- $\nu_\gamma = 0.777$

$$\gamma^\star = \frac{(1 - \sqrt{1-\nu})^2}{\nu}$$

$$\nu^\star = 1 - \sqrt{1-\nu}$$

- Depends on the *largest* eigenvalue
- Worst case at rate ν : ν
- In this example $\nu = 0.85$

$$\begin{cases} y_{k+1} = Rx^k + d \\ x_{k+1} = y_{k+1} + \gamma(y_{k+1} - y_k) \end{cases} \Rightarrow \mathbf{R}^\gamma = \begin{bmatrix} (1+\gamma)R & -\gamma R \\ I & 0 \end{bmatrix} \text{ on } \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix}$$



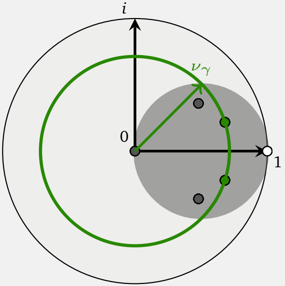
eigenvalues of R^γ

- $\gamma = 0.442 = \gamma^\star$
- $\nu_\gamma = 0.613 = \nu^\star$

$$\gamma^\star = \frac{(1-\sqrt{1-\nu})^2}{\nu}$$
$$\nu^\star = 1 - \sqrt{1-\nu}$$

- Depends on the *largest* eigenvalue
- Worst case at rate ν : ν
- In this example $\nu = 0.85$

$$\begin{cases} y_{k+1} = Rx^k + d \\ x_{k+1} = y_{k+1} + \gamma(y_{k+1} - y_k) \end{cases} \Rightarrow \mathbf{R}^\gamma = \begin{bmatrix} (1+\gamma)R & -\gamma R \\ I & 0 \end{bmatrix} \text{ on } \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix}$$



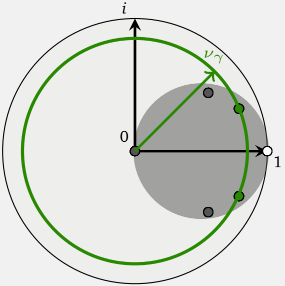
eigenvalues of R^γ

- $\gamma = 0.6$
- $\nu_\gamma = 0.714$

$$\gamma^\star = \frac{(1-\sqrt{1-\nu})^2}{\nu}$$
$$\nu^\star = 1 - \sqrt{1-\nu}$$

- Depends on the *largest* eigenvalue
- Worst case at rate ν : ν
- In this example $\nu = 0.85$

$$\begin{cases} y_{k+1} = Rx^k + d \\ x_{k+1} = y_{k+1} + \gamma(y_{k+1} - y_k) \end{cases} \Rightarrow \mathbf{R}^\gamma = \begin{bmatrix} (1+\gamma)R & -\gamma R \\ I & 0 \end{bmatrix} \text{ on } \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix}$$



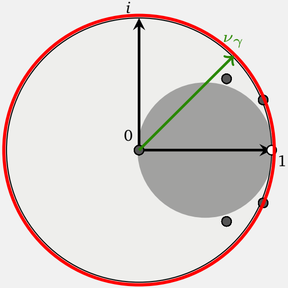
eigenvalues of R^γ

- $\gamma = 0.85$
- $\nu_\gamma = 0.85$

$$\gamma^\star = \frac{(1-\sqrt{1-\nu})^2}{\nu}$$
$$\nu^\star = 1 - \sqrt{1-\nu}$$

- Depends on the *largest* eigenvalue
- Worst case at rate ν : ν
- In this example $\nu = 0.85$

$$\begin{cases} y_{k+1} = Rx^k + d \\ x_{k+1} = y_{k+1} + \gamma(y_{k+1} - y_k) \end{cases} \Rightarrow \mathbf{R}^\gamma = \begin{bmatrix} (1+\gamma)R & -\gamma R \\ I & 0 \end{bmatrix} \text{ on } \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix}$$



eigenvalues of R^γ

- $\gamma = 1.2$
- $\nu_\gamma = 1.01$

$$\gamma^\star = \frac{(1-\sqrt{1-\nu})^2}{\nu}$$
$$\nu^\star = 1 - \sqrt{1-\nu}$$

- Depends on the *largest* eigenvalue
- Worst case at rate $\nu : \nu$
- In this example $\nu = 0.85$

Online Inertia for a FNE operator $\mathcal{M}$:

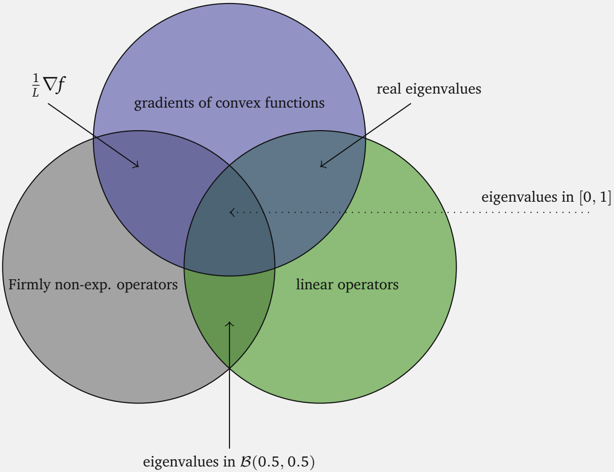
[rate estimation] $$v_k = \sqrt{\frac{\|x_k - x_{k-1}\|^2 + \|x_{k-1} - x_{k-2}\|^2}{\|x_{k-1} - x_{k-2}\|^2 + \|x_{k-2} - x_{k-3}\|^2}}$$

[virtual max. eigenvalue] $$\boldsymbol{\nu}_k = Proj_{[\varepsilon, 1-\varepsilon]}\left(\frac{(v_k)^2}{\gamma_k v_k - \gamma_k + v_k}\right)$$

[deduced opt. paramater] $$\gamma_{k+1} = \gamma_{k+2} = \frac{(1 - \sqrt{1 - \boldsymbol{\nu}_k})^2}{\boldsymbol{\nu}_k}$$

$$y_{k+1} = \mathcal{M}(x_k) \quad x_{k+1} = y_{k+1} + \gamma_{k+1}(y_{k+1} - y_k)$$
$$y_{k+2} = \mathcal{M}(x_{k+1}) \quad x_{k+2} = y_{k+2} + \gamma_{k+2}(y_{k+2} - y_{k+1})$$

- same intuition
- convergence ensured by **restart** as $\gamma_k \in [0, 1[$
- no monotonicity



- every subdifferential of a convex function is a monotone operator
- every cyclically monotone operator is a subdifferential [Rockafellar'67]
- cyclically monotone linear operator have real eigenvalues [Shiu'76]

– worst case for relaxation in the intersection, not for inertia

– ADMM can be casted as a gradient descent for some functions [Patrinos et al.'14]

Acceleration & Operators

Relaxation Inertia

in Practice

Relaxation Inertia **Application**

Bridging Relaxation & Inertia

Intuition Alt. Inertia

the Proximal Gradient algorithm

Acceleration Alt. inertia

- We have efficient methods to choose relaxation or inertia parameter...
- ...based on the contraction verified by hyper-parameter $\zeta_k = \rho z_k + \lambda_k$

  Problem: the mapping $\zeta \leftrightarrow (z, \lambda)$ is *non-linear*

Relaxed ADMM

$$x_{k+1} = \arg\min_x \left\{ f(x) + \frac{\rho}{2} \left\| Mx - z_k + \frac{\lambda_k}{\rho} \right\|^2 \right\}$$

$$z_{k+1} = \arg\min_z \left\{ g(z) + \frac{\rho}{2} \left\| Mx_{k+1} - z + \frac{\lambda_k}{\rho} + (\eta_k - 1)(Mx_{k+1} - z_k) \right\|^2 \right\}$$

$$\lambda_{k+1} = \lambda_k + \rho(Mx_{k+1} - z_{k+1} + (\eta_k - 1)(Mx_{k+1} - z_k))$$

– obtained by monotone operator *representation* lemma (see e.g. [Eckstein'92])

- We have efficient methods to choose relaxation or inertia parameter...
- ...based on the contraction verified by hyper-parameter $\zeta_k = \rho z_k + \lambda_k$

Problem: the mapping $\zeta \leftrightarrow (z, \lambda)$ is *non-linear*

Inertial ADMM

$$x_{k+1} = \arg\min_x \left\{ f(x) + \frac{\rho}{2} \left\| Mx - z_k + \frac{\lambda_k}{\rho} \right\|^2 \right\}$$

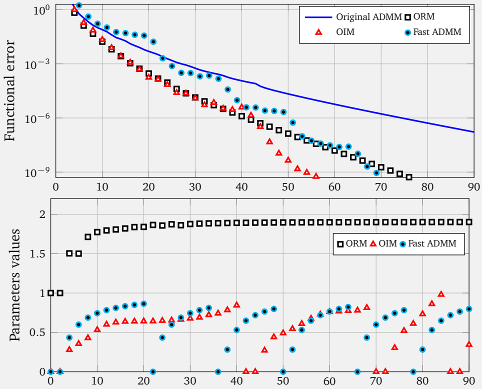
$$z_{k+1} = \arg\min_z \left\{ g(z) + \frac{\rho}{2} \left\| Mx_{k+1} - z + \frac{\lambda_k}{\rho} + \gamma_k \left(M(x_{k+1} - x_k) + \frac{\lambda_k - \lambda_{k-1}}{\rho} \right) \right\|^2 \right\}$$

$$\lambda_{k+1} = \lambda_k + \rho \left(Mx_{k+1} - z_{k+1} + \gamma_k \left(M(x_{k+1} - x_k) + \frac{\lambda_k - \lambda_{k-1}}{\rho} \right) \right)$$

– also obtained by monotone operator *representation* lemma

– **different** from *Fast ADMM* [Golstein et al.'14] except for indicators and quadratics

lasso problem: $\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1$ (300 × 100) 10% sparsity



- Online Relaxation is steady in acceleration and parameters
- Online Inertia is more careful than Fast ADMM and thus restarts less leading to better performance

- Relaxation and Inertia do not mix well...
- Reasoning can be extended to general α-averaged operators

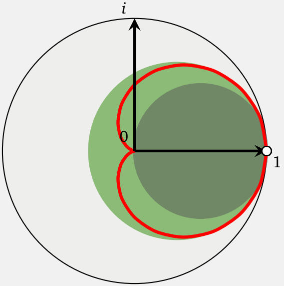
$$\|\mathcal{M}(x) - x^\star\|^2 \leq \|x - x^\star\|^2 - \frac{1-\alpha}{\alpha}\|\mathcal{M}(x) - x\|^2 \quad \alpha \in]0,1[$$

$\alpha = \frac{1}{2}$ is the previous *Firm non-expansiveness*

Proximal gradient: $\mathcal{M}_{prox.\ grad.} = \underbrace{\underbrace{\mathcal{M}_{prox.}}_{\alpha=1/2} \circ \underbrace{\mathcal{M}_{grad.}}_{\alpha=1/2}}_{\alpha=2/3}$

but...

gray: $\alpha = 1/2$
green: $\alpha = 2/3$
red: Composition of two $\alpha = 1/2$



eigenvalues of R

Acceleration & Operators

in Practice

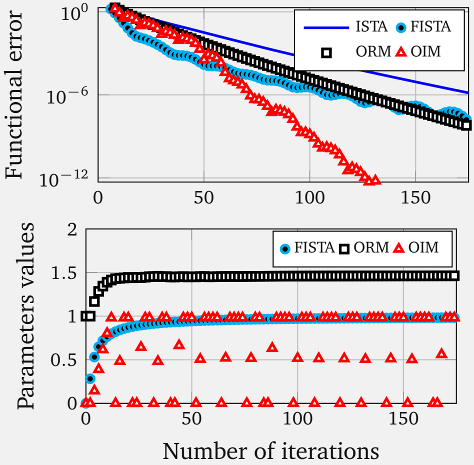
■ Bridging Relaxation & Inertia

the Proximal Gradient algorithm

- Online acceleration methods

Relaxation:	+ stability	− acceleration
Inertia:	− stability (restart)	+ acceleration

lasso
Proximal Gradient



Acceleration & Operators

Relaxation   Inertia

in Practice

Relaxation   Inertia   Application

Bridging Relaxation & Inertia

Intuition   **Alt. Inertia**

the Proximal Gradient algorithm

Acceleration   Alt. inertia

$$\begin{cases} y_{k+1} = \mathcal{M}(x_k) \\ y_{k+2} = \mathcal{M}(y_{k+1}) \\ \mathbf{x_{k+2}} = \mathbf{y_{k+2}} + \gamma_{\mathbf{k+2}}(\mathbf{y_{k+2}} - \mathbf{y_{k+1}}) \end{cases}$$

with $\mathcal{M}$ firmly non-expansive

Alternated Inertia *converges if* $0 \leq \gamma_k \leq 1$

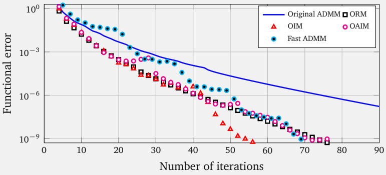
- Fejér monotonous at least with this condition
- possibly converging under broader conditions
- introduced in [Mu'15;I.-Hendrickx'16]

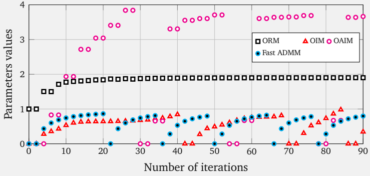
in Practice:

- one can also choose Nesterov's sequence or even 1...
- but the same eigenvalue-based analysis can be conducted
  → **Online Alternated Inertia Method (OAIM)**

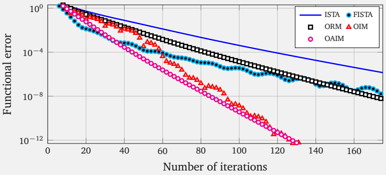
$$\gamma^\star = \frac{2\nu^2 + (\sqrt{2}-1)\nu}{2\nu(1-\nu) + 1/2} \qquad \nu^\star = \frac{\gamma^\star}{2\sqrt{1+\gamma^\star}}$$

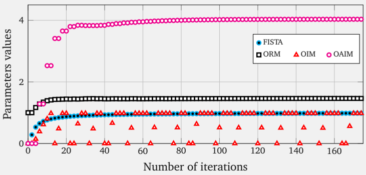
ADMM

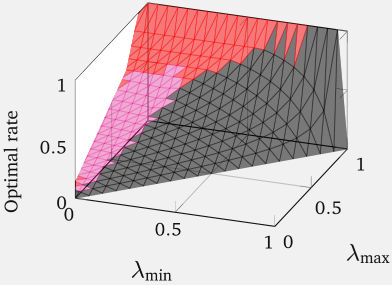




Proximal gradient



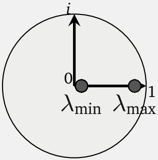




▶ If $\lambda_{\min} = 0$ "good stepsize", Alternated In. better than In. if $\lambda_{\max} \leq 1 - \underbrace{\left(\frac{4}{9+4\sqrt{2}}\right)}_{\mu/L \approx 0.273}$

▶ If $\lambda_{\min} >> 0$ "bad stepsize", Relaxation is better for well-conditioned problems.

Best rate for a linear operator with real eig.



attained by

Relaxation
Inertia
Alternated Inertia

Example: $f(x) = ||Ax - b||_2^2$
gradient operator $\mathcal{M}(x) = (I - \gamma(2A^{\mathrm{T}}A))x + 2A^{\mathrm{T}}b$
$\lambda_{\min} = 1 - \gamma L$, $\lambda_{\max} = 1 - \gamma\mu$

When the rate is sublinear ($\mathcal{O}(1/k), \mathcal{O}(1/k^2)$), popular parameters choice are

Relaxation	**Inertia**	**Alternated Inertia**	
$\eta \to 2$	$\gamma \to 1$	$\gamma \to 1$	$\gamma \to 2 + 2\sqrt{2}$

but if some *small undetected* strong convexity $\mu/L > 0$ is present,
the limit **linear rate** for a linear sym. FNE operator is

$1 - 2\frac{\mu}{L}$	**1**	$1 - \frac{3}{2}\frac{\mu}{L}$	$1 - \left(2 + \frac{3}{\sqrt{2}}\right)\frac{\mu}{L}$

- Practical interest of Alternated Inertia
- *Functional* analysis in the case of the Proximal Gradient

Acceleration & Operators

in Practice

Bridging Relaxation & Inertia

- the Proximal Gradient algorithm

Acceleration & Operators

Relaxation Inertia

in Practice

Relaxation Inertia Application

Bridging Relaxation & Inertia

Intuition Alt. Inertia

the Proximal Gradient algorithm

Acceleration Alt. inertia

Problem $\min_x F(x) := f(x) + g(x)$ with f smooth

Proximal gradient operator for $F := f + g$ and step α:
$\mathsf{T}_\alpha(x) = \mathbf{prox}_{\alpha g}\left(x - \alpha \nabla f(x)\right)$.

Acceleration *via* extrapolation: $\begin{cases} y_{k+1} = \mathsf{T}_\alpha(x_k) \\ x_{k+1} = \textbf{extrapolation}\left(\{y_\ell\}_{\ell \leq k+1}\right) \end{cases}$

extrapolation is **typically** a linear combination $x_{k+1} = y_{k+1} + \gamma_k(y_{k+1} - y_k)$
based on coefficients of the type [Nesterov'83;Aujol-Dossal'15]

$$\gamma_k = \frac{t_k - 1}{t_{k+1}} \qquad \to 1 \text{ at rate } \frac{1}{k^d}, d \in (0, 1]$$

$$t_0 = 0 \text{ and } t_k := \left(\frac{k + a - 1}{a}\right)^d \text{ or } \frac{1 + \sqrt{1 + 4t_{k-1}^2}}{2}$$

FISTA: $\begin{cases} y_{k+1} = \mathsf{T}_\alpha(x_k) \\ x_{k+1} = y_{k+1} + \gamma_{k+1}(y_{k+1} - y_k) \end{cases}$ with $\gamma_{k+1} = \frac{t_k - 1}{t_{k+1}}$; $t_k = \frac{k+a+1}{a}$ or $\frac{1+\sqrt{1+4t_{k-1}^2}}{2}$.

with $\alpha = \frac{1}{L}$,

$$\begin{aligned} t_{k+1}^2 F(y_{k+2}) - t_k^2 F(y_{k+1}) \\ \leq -\frac{1}{2\gamma} \|t_{k+1} y_{k+2} - (t_{k+1} - 1) y_{k+1} - y^\star\|^2 \\ + \frac{1}{2\gamma} \|t_{k+1} x_{k+1} - (t_{k+1} - 1) y_{k+1} - y^\star\|^2 \end{aligned}$$

$$\begin{aligned} t_k^2 F(y_{k+1}) - t_{k-1}^2 F(y_k) \\ \leq -\frac{1}{2\gamma} \|t_k y_{k+1} - (t_k - 1) y_k - y^\star\|^2 \\ + \frac{1}{2\gamma} \|t_k x_k - (t_k - 1) y_k - y^\star\|^2 \end{aligned}$$

telescoping if $x_{k+1} = y_{k+1} + \frac{t_k - 1}{t_{k+1}}(y_{k+1} - y_k)$

Rate $t_k^2 F(y_{k+1}) \leq C$ thus $F(y_{k+1}) \leq \frac{C}{t_k^2} = \mathcal{O}\left(\frac{1}{k^2}\right)$

Acceleration & Operators

Relaxation Inertia

in Practice

Relaxation Inertia Application

Bridging Relaxation & Inertia

Intuition Alt. Inertia

the Proximal Gradient algorithm

Acceleration **Alt. inertia**

Acceleration alternated extrapolation:

$$\begin{cases} x_k = y_k & x_{k+1} = \textbf{extrapolation}\left(\{y_\ell\}_{\ell \leq k+1}\right) \\ y_{k+1} = \mathsf{T}_\alpha(x_k) & y_{k+2} = \mathsf{T}_\alpha(x_{k+1}) \end{cases}$$

Choice 1: $1/k^2$ rate $\quad x_{k+1} = y_{k+1} - \frac{1}{t_{k+1}}(y_{k+1} - y_k) + \frac{t_k - 1}{t_{k+1}}(y_k - y_{k-1})$

with $t_k = \frac{k+a+1}{a}$ or $\frac{1+\sqrt{1+4t_{k-1}^2}}{2}$ and $\alpha = \frac{1}{L}$

$$F(y_{k+2}) = \mathcal{O}\left(\frac{1}{k^2}\right)$$

- $F(y_{2k})$ is *non-monotonous*
- Alternated **Heavy balls**

Choice 2: alternated inertia $\quad x_{k+1} = y_{k+1} + \gamma_{k+1}(y_{k+1} - y_k)$

$$F(y_{k+2}) \leq F(y_k) - \frac{(2 - \alpha L - \gamma_{k+1})}{2}\left(\|y_{k+1} - x_k\|^2 + \|y_{k+2} - x_{k+1}\|^2\right)$$

- $F(y_{2k})$ is *non-increasing* for $\alpha = 1/L$ and $\gamma_k \in [0, 1]$
- Rate???

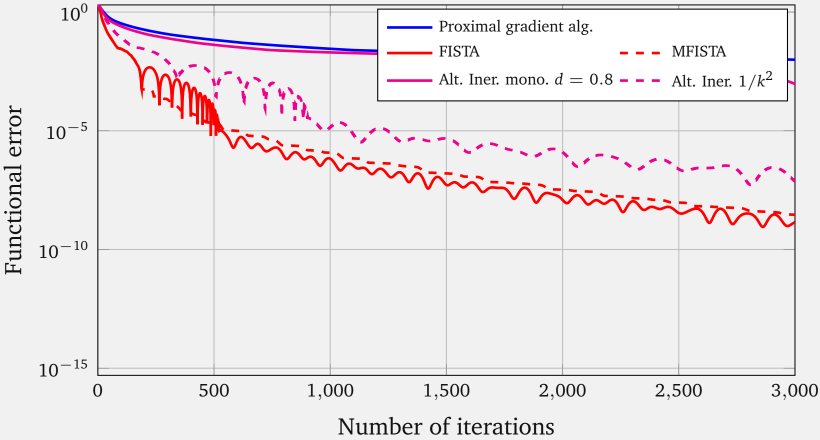
- F is a KL function with $(F(u)-F^\star)^{1-\theta} \leq C.dist(0,\partial F(u))$ for all $u : F(u) < F^\star + \eta$ some $C, \eta > 0$, $\theta \in (0,1]$
- $\mathcal{M}$ produce (x_k) such that

$$F(x_{k+1}) \leq F(x_k) - a_k[\text{dist}(0,\partial F(x_{k+1}))]^2 \quad \text{with} \quad a_k > 0 \text{ and } \sum_{k=1}^{\infty} a_k = +\infty$$

Alt. Iner. for PG: $F(y_{k+2}) \leq F(y_k) - \frac{(2-\alpha L-\gamma_{k+1})}{2}(\|y_{k+1}-x_k\|^2 + \|y_{k+2}-x_{k+1}\|^2)$

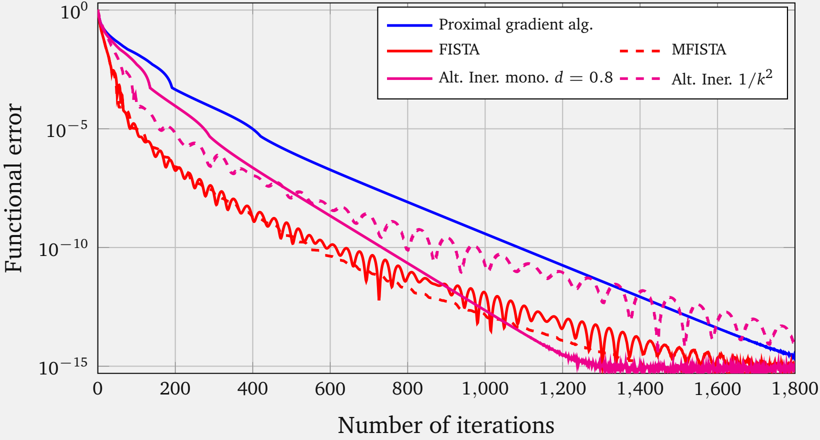
$\theta = 1$			finite number of steps
$\theta \in [0.5, 1[$	$a_k \geq a > 0$	$\gamma_k \leq \gamma < 1$ or stepsize $< 1/L$	$\mathcal{O}\left(\left[\frac{C^2}{C^2+1}\right]^k\right)$
	$a_k = \frac{1}{k}$	Nesterov, $d = 1$	$\mathcal{O}\left(\frac{1}{k^{\frac{1}{2C^2}}}\right)$
	$a_k = \frac{1}{k^d}, d \in]0,1[$	$d \in]0,1[$	$\mathcal{O}\left(\exp\left(-\frac{k^d}{2C^2}\right)\right)$
$\theta \in]0, 0.5[$	$a_k \geq a > 0$	$\gamma_k \leq \gamma < 1$ or stepsize $< 1/L$	$\mathcal{O}\left(\frac{1}{k^{1+\frac{2\theta}{1-2\theta}}}\right)$
	$a_k = \frac{1}{k}$	Nesterov, $d = 1$	$\mathcal{O}\left(\frac{1}{\log(k)^{1+\frac{2\theta}{1-2\theta}}}\right)$
	$a_k = \frac{1}{k^d}, d \in]0,1[$	$d \in]0,1[$	$\mathcal{O}\left(\frac{1}{k^{1+\frac{2\theta-1+d}{1-2\theta}}}\right)$

ℓ_1 regularized logisitic regression. `ionosphere` dataset (351×35) 50% sparsity



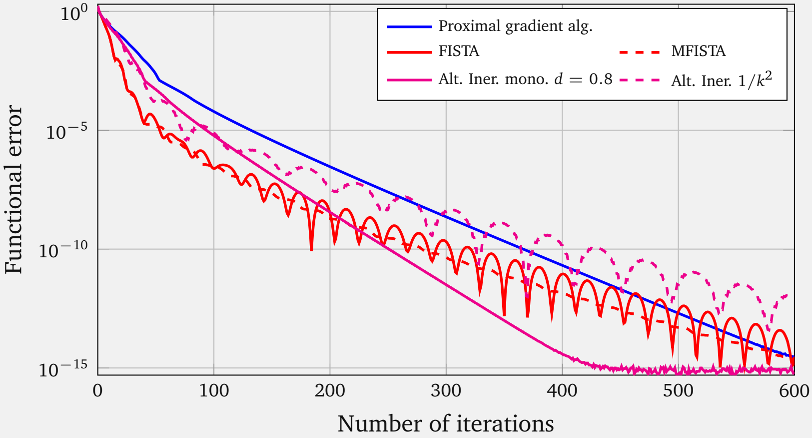
$1/L_{upper\ bound}$ pessimistic stepsize

ℓ_1 regularized logisitic regression. `ionosphere` dataset (351×35) 50% sparsity



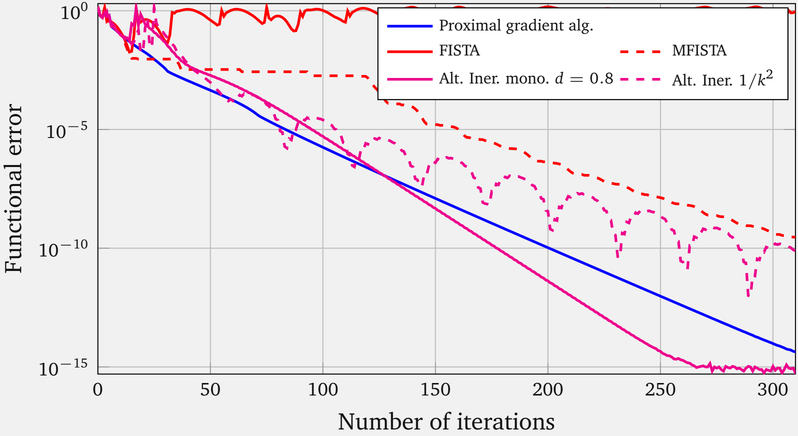
$\alpha = 8$ times less than the maximal stepsize for PG

ℓ_1 regularized logisitic regression. `ionosphere` dataset (351×35) 50% sparsity



$\alpha = 3$ times less than the maximal stepsize for PG

ℓ_1 regularized logisitic regression. `ionosphere` dataset (351×35) 50% sparsity



$\alpha = 1.5$ times less than the maximal stepsize for PG

Practical Acceleration of various algorithms:

- Methods to very simply accelerate a class of optimization methods
- Relaxation is more stable; Inertia can be more efficient
- Alternated Inertia can be a compromise

Limitations and Perspectives:

- Are complex methods “gradient-like” ?
- Speed/stability tradeoff without restart?

I did not talk about:

- Restart [Fercoq-Qu’16;Roulet-d’Aspremont’16]
- More complex methods [Scieur-Roulet-Bach-d’Aspremont’17;next talks]
- Non-convexity