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 $x^*$

Optimization viewpoint:

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

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

towards minimizers $F^*$

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 $x^*$ and any $x$

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

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 $\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2$
- $O(1/k)$ 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;...]
Extrapolation steps

\[
\begin{align*}
    y_{k+1} &= \mathcal{M}(x_k) \\
    x_{k+1} &= y_{k+1} \text{ extrapolation} (y_{k+1}, (y_k), (x_k))
\end{align*}
\]

Assumption:
The fixed point method \( \mathcal{M} \) is firmly non-expansive i.e. for any fixed point \( x^* \) and any \( x \)

\[
    \| \mathcal{M}(x) - x^* \|^2 \leq \| x - x^* \|^2 - \| \mathcal{M}(x) - x \|^2.
\]

Acceleration:
- operation output \( y_{k+1} \)

Using
- past outputs \( y_k, y_{k-1}, \ldots \)
- past iterates \( x_k, x_{k-1}, \ldots \)

Two main strategies:
- Relaxation \( x_{k+1} = y_{k+1} + (\eta_k - 1)(y_{k+1} - x_k) \)
  plays on the methods contraction.
- Inertia \( x_{k+1} = y_{k+1} + \gamma_k (y_{k+1} - y_k) \)
  plays on the moments of the iterates sequence.
ACCELERATION & OPERATORS
Relaxation Inertia

IN PRACTICE
Relaxation Inertia Application

BRIDGING RELAXATION & INERTIA
Intuition Alt. Inertia

THE PROXIMAL GRADIENT ALGORITHM
Acceleration Alt. inertia
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 \( \eta \)
- Optimal \( \eta \) gives Chebyshev iterations

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

Krasnoselskiǐ–Mann iterations (1955): Relaxation is present in the operator convergence theorem.
\[
\begin{align*}
    y_{k+1} &= M(x_k) \\
    x_{k+1} &= y_{k+1} + (\eta_{k+1} - 1)(y_{k+1} - x_k)
\end{align*}
\] with \( M \) firmly non-expansive

**Relaxation converges if** \( 0 < \lim \inf \eta_k \leq \lim \sup \eta_k < 2 \).

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

**Gradient algorithm:**

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

- “optimal” \( \frac{2}{1+\mu/L} \) with \( \mu \)-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 \( \eta_k \) 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 $f$

\[
\begin{align*}
    y_{k+1} &= x_k - \frac{1}{L} \nabla f(x_k) \\
    x_{k+1} &= y_{k+1} + \gamma_{k+1} (y_{k+1} - y_k)
\end{align*}
\]

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

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

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

- Faster (sub-linear) convergence rate: $O(1/k) \to O(1/k^2)$
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: $\ddot{x}(t) + \alpha(t)\dot{x}(t) = -\nabla f(x(t))$

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

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

- $\alpha(t) = \alpha \rightarrow$ fixed inertia; $\alpha(t) = \alpha/t \rightarrow \gamma_k = \frac{k-1}{k+\alpha-1}$.

- Used recently [Attouch’15] to prove iterates convergence of accelerated Forward-Backward

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

\begin{equation*}
\begin{cases}
y_{k+1} = \mathcal{M}(x_k) \\
x_{k+1} = y_{k+1} + \gamma_{k+1}(y_{k+1} - y_k)
\end{cases}
\end{equation*}

with \(\mathcal{M}\) firmly non-expansive

**Inertia converges if** \(\lim \sup \gamma_k < 1/3\)

- 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{equation*}
\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}
\end{equation*}

- “optimal” \(\frac{1 - \sqrt{\mu/L}}{1 + \sqrt{\mu/L}}\) with \(\mu\)-strong convexity

**ADMM:**

- Update is more involved (see later)

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

- [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 \( \| \mathcal{M}(x) - x^* \|^2 \leq \| x - x^* \|^2 - \| \mathcal{M}(x) - x \|^2 \)

- *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* \( \nu \)
  
  \( \| x_k - x^* \| = \tilde{O}(\nu^k) \)

- 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
Effect of Relaxation on $R$

\[
\begin{align*}
    y_{k+1} &= Rx_k + d \\
    x_{k+1} &= y_{k+1} + (\eta - 1)(y_{k+1} - x_k)
\end{align*}
\]

\[\Rightarrow R_\eta = \eta R + (1 - \eta)I \quad \text{on } x_k\]

Eigenvalues of $R_\eta$

\[\eta^* = \frac{2}{2 - \nu}\]
\[\nu^* = \frac{\nu}{2 - \nu}\]

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

$\eta = 1$

$\nu_\eta = 0.75$
Effect of Relaxation on $R$

$$\begin{align*}
y_{k+1} &= Rx_k + d \\
x_{k+1} &= y_{k+1} + (\eta - 1)(y_{k+1} - x_k)
\end{align*}$$

$\Rightarrow R_\eta = \eta R + (1 - \eta)I$ on $x_k$

$\nu_\eta = 0.82$

$\eta = 0.7$

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

$$\begin{align*}
\eta^* &= \frac{2}{2 - \nu} \\
\nu^* &= \frac{\nu}{2 - \nu}
\end{align*}$$

eigenvalues of $R_\eta$
Effect of Relaxation on $R$

\[
\begin{align*}
  y_{k+1} &= Rx_k + d \\
  x_{k+1} &= y_{k+1} + (\eta - 1)(y_{k+1} - x_k)
\end{align*}
\]

$\Rightarrow R_\eta = \eta R + (1 - \eta)I$ on $x_k$

\[\begin{aligned}
\eta^* &= \frac{2}{2 - \nu} \\
\nu^* &= \frac{\nu}{2 - \nu}
\end{aligned}\]

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

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

Eigenvalues of $R_\eta$
Effect of Relaxation on $R$

\[
\begin{align*}
    y_{k+1} &= Rx_k + d \\
    x_{k+1} &= y_{k+1} + (\eta - 1)(y_{k+1} - x_k)
\end{align*}
\]

$\Rightarrow R_\eta = \eta R + (1 - \eta)I$ on $x_k$

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

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

$\nu^* = \frac{\nu}{2 - \nu}$

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

\[
\begin{align*}
    y_{k+1} &= Rx_k + d \\
    x_{k+1} &= y_{k+1} + (\eta - 1)(y_{k+1} - x_k)
\end{align*}
\]

$\Rightarrow R_\eta = \eta R + (1 - \eta)I \quad \text{on } x_k$

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

Depends on *extremal* eigenvalues

- Worst case at rate $\nu : [0, \nu]$
- In this example $\nu = 0.75$

\[\eta^* = \frac{2}{2 - \nu}\]
\[\nu^* = \frac{\nu}{2 - \nu}\]
Effect of Relaxation on $R$

\[
\begin{aligned}
    y_{k+1} &= Rx_k + d \\
    x_{k+1} &= y_{k+1} + (\eta - 1)(y_{k+1} - x_k)
\end{aligned}
\Rightarrow R_\eta = \eta R + (1 - \eta)I \text{ on } x_k
\]

\[\eta^* = \frac{2}{2 - \nu}\]
\[\nu^* = \frac{\nu}{2 - \nu}\]

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

$\eta = 1.9$
$\nu_\eta = 0.9$
Effect of Relaxation on $R$

\[
\begin{align*}
    y_{k+1} &= R x_k + d \\
    x_{k+1} &= y_{k+1} + (\eta - 1)(y_{k+1} - x_k)
\end{align*}
\Rightarrow R_\eta = \eta R + (1 - \eta)I \quad \text{on } x_k
\]

\[\eta^* = \frac{2}{2 - \nu}\]
\[\nu^* = \frac{\nu}{2 - \nu}\]

- 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 $v_k = \eta_k \nu_k + (1 - \eta_k) \Rightarrow \nu_k = \frac{v_k - 1 + \eta_k}{\eta_k}$

3. Optimal relaxation on $\nu_k$, $\eta_{k+1} = \frac{2}{2 - \nu_k} = \frac{2\eta_k}{\eta_{k+1} - \nu_k}$

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

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

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

- $\nu_k$ is simplistic but theoretically consistent rate approx. as $\nu_k \in [0, 1]$
- we prove that $\eta_k \in \left[\frac{\varepsilon}{2}; 2 - \frac{\varepsilon}{2}\right]$ 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
Effect of Inertia on $R$

\[
\begin{align*}
  y_{k+1} &= Rx^k + d \\
  x_{k+1} &= y_{k+1} + \gamma(y_{k+1} - y_k)
\end{align*}
\Rightarrow \quad R^\gamma = \begin{bmatrix}
  (1 + \gamma)R & -\gamma R \\
  I & 0
\end{bmatrix}
\text{ on } \begin{bmatrix}
  x_k \\
  x_{k-1}
\end{bmatrix}
\]

\[\nu_{\gamma} = 0.85\]

\[\nu^* = 1 - \sqrt{1 - \nu}\]

\[\gamma^* = \frac{(1 - \sqrt{1 - \nu})^2}{\nu}\]

\[\gamma = 0\]

- Depends on the largest eigenvalue
- Worst case at rate $\nu : \nu$
- In this example $\nu = 0.85$
\[
\begin{align*}
\begin{cases}
y_{k+1} &= Rx^k + d \\
x_{k+1} &= y_{k+1} + \gamma(y_{k+1} - y_k)
\end{cases} \Rightarrow R^\gamma &= \begin{bmatrix}
(1 + \gamma)R & -\gamma R \\
I & 0
\end{bmatrix}
on \begin{bmatrix}
x_k \\
x_{k-1}
\end{bmatrix}
\end{align*}
\]
Effect of Inertia on $R$

\[
\begin{align*}
y_{k+1} &= Rx_k^k + d \\
x_{k+1} &= y_{k+1} + \gamma(y_{k+1} - y_k)
\end{align*}
\]

\[\Rightarrow R^\gamma = \begin{bmatrix} (1 + \gamma)R & -\gamma R \\ I & 0 \end{bmatrix} \text{ on } \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix}\]

\[
\nu^\gamma = 0.777
\]

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

- Depends on the largest eigenvalue
- Worst case at rate $\nu : \nu$
- In this example $\nu = 0.85$
Effect of Inertia on \( R \)

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

eigenvalues of \( R^\gamma \)

\[
\gamma^* = \frac{(1 - \sqrt{1 - \nu})^2}{\nu}
\]

\[
\nu^* = 1 - \sqrt{1 - \nu}
\]

- \( \gamma = 0.442 = \gamma^* \)
- \( \nu_\gamma = 0.613 = \nu^* \)
- Depends on the largest eigenvalue
- Worst case at rate \( \nu : \nu \)
- In this example \( \nu = 0.85 \)
Effect of Inertia on $R$

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

Eigenvalues of $R^\gamma$

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

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

- Depends on the largest eigenvalue
- Worst case at rate $\nu : \nu$
- In this example $\nu = 0.85$
Effect of Inertia on $R$

\[
\begin{align*}
    y_{k+1} &= Rx^k + d \\
    x_{k+1} &= y_{k+1} + \gamma(y_{k+1} - y_k)
\end{align*}
\]

\[R^\gamma = \begin{bmatrix}
    (1 + \gamma)R & -\gamma R \\
    I & 0
\end{bmatrix}
\]

on \[
\begin{bmatrix}
    x_k \\
    x_{k-1}
\end{bmatrix}
\]

\[\gamma^* = \frac{(1 - \sqrt{1 - \nu})^2}{\nu}
\]

\[\nu^* = 1 - \sqrt{1 - \nu}
\]

- Depends on the largest eigenvalue
- Worst case at rate $\nu : \nu$
- In this example $\nu = 0.85$
Effect of Inertia on $R$

\[
\begin{align*}
    y_{k+1} &= R x^k + d \\
    x_{k+1} &= y_{k+1} + \gamma (y_{k+1} - y_k)
\end{align*}
\Rightarrow \quad R^\gamma = \begin{bmatrix} (1 + \gamma)R & -\gamma R \\ I & 0 \end{bmatrix} \text{ on } \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix}
\]

\[
\begin{align*}
    \nu_\gamma &= 1.01 \\
    \gamma &= 1.2
\end{align*}
\]

\[
\begin{align*}
    \gamma^* &= \frac{(1 - \sqrt{1 - \nu})^2}{\nu} \\
    \nu^* &= 1 - \sqrt{1 - \nu}
\end{align*}
\]

- 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] $\nu_k = \text{Proj}_{[\epsilon, 1-\epsilon]} \left( \frac{(v_k)^2}{\gamma_k v_k - \gamma_k + v_k} \right)$

- [deduced opt. parameter] $\gamma_{k+1} = \gamma_{k+2} = \frac{(1 - \sqrt{1 - \nu_k})^2}{\nu_k}$

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

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

- same intuition
- convergence ensured by restart as $\gamma_k \in [0, 1]$
- no monotonicity
Why is the theoretical limit $\frac{1}{3}$ but greater values are used?  

**Accel. in Practice**

- 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**

\[
\begin{align*}
x_{k+1} &= \arg \min_x \left\{ f(x) + \frac{\rho}{2} \left\| Mx - z_k + \frac{\lambda_k}{\rho} \right\|^2 \right\} \\
z_{k+1} &= \arg \min_z \left\{ g(z) + \frac{\rho}{2} \left\| Mx_{k+1} - z + \frac{\lambda_k}{\rho} + (\eta_k - 1)(Mx_{k+1} - z_k) \right\|^2 \right\} \\
\lambda_{k+1} &= \lambda_k + \rho (Mx_{k+1} - z_{k+1} + (\eta_k - 1)(Mx_{k+1} - z_k))
\end{align*}
\]

-- 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

\[
\begin{align*}
x_{k+1} &= \arg \min_x \left\{ f(x) + \frac{\rho}{2} \left\| Mx - z_k + \frac{\lambda_k}{\rho} \right\|^2 \right\} \\
z_{k+1} &= \arg \min_z \left\{ g(z) + \frac{\rho}{2} \left\| Mx_{k+1} - z + \frac{\lambda_k}{\rho} + \gamma_k \left( M(x_{k+1} - x_k) + \frac{\lambda_k - \lambda_{k-1}}{\rho} \right) \right\|^2 \right\} \\
\lambda_{k+1} &= \lambda_k + \rho \left( Mx_{k+1} - z_{k+1} + \gamma_k \left( M(x_{k+1} - x_k) + \frac{\lambda_k - \lambda_{k-1}}{\rho} \right) \right)
\end{align*}
\]

– 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
Reach and limits of the approach

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

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

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

**Proximal gradient:** $M_{\text{prox. grad.}} = M_{\text{prox.}} \circ M_{\text{grad.}}$

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

**Diagram:**
- Eigenvalues of $R$
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
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
Alternated Inertia

\[
\begin{aligned}
  y_{k+1} &= \mathcal{M}(x_k) \\
  y_{k+2} &= \mathcal{M}(y_{k+1}) \\
  x_{k+2} &= y_{k+2} + \gamma_{k+2}(y_{k+2} - y_{k+1})
\end{aligned}
\]

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)

\[
\gamma^* = \frac{2\nu^2 + (\sqrt{2} - 1)\nu}{2\nu(1 - \nu) + 1/2} \quad \nu^* = \frac{\gamma^*}{2\sqrt{1 + \gamma^*}}
\]
Illustration on the lasso problem

**ADMM**

![ADMM Diagram](Image)

**Proximal gradient**

![Proximal gradient Diagram](Image)
Comparison 1 – Quadratic rates

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

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

Best rate for a linear operator with real eig.

Example: $f(x) = \|Ax - b\|_2^2$

gradient operator $M(x) = (I - \gamma(2A^TA))x + 2A^Tb$

$\lambda_{\min} = 1 - \gamma L$, $\lambda_{\max} = 1 - \gamma \mu$
When the rate is sublinear ($\mathcal{O}(1/k), \mathcal{O}(1/k^2)$), popular parameters choice are

<table>
<thead>
<tr>
<th>Relaxation</th>
<th>Inertia</th>
<th>Alternated Inertia</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\eta \rightarrow 2$</td>
<td>$\gamma \rightarrow 1$</td>
<td>$\gamma \rightarrow 2 + 2\sqrt{2}$</td>
</tr>
</tbody>
</table>

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

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

- **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 \( \alpha \):
\[
T_\alpha(x) = \text{prox}_{\alpha g} (x - \alpha \nabla f(x)).
\]

**Acceleration via extrapolation:**
\[
\left\{ \begin{array}{l}
y_{k+1} = T_\alpha(x_k) \\
x_{k+1} = \text{extrapolation} (\{y_\ell\}_{\ell \leq k+1})
\end{array} \right.
\]

*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 \([\text{Nesterov’83; Aujol-Dossal’15}]\)
\[
\gamma_k = \frac{t_k - 1}{t_{k+1}} \quad \rightarrow 1 \text{ at rate } \frac{1}{k^d}, \; d \in (0, 1]
\]

\[
t_0 = 0 \text{ and } t_k := \left( \frac{k + a - 1}{a} \right)^d \text{ or } \frac{1 + \sqrt{1 + 4t_{k-1}^2}}{2}
\]
FISTA: \[
\begin{align*}
    y_{k+1} &= \mathcal{T}_\alpha(x_k) \\
    x_{k+1} &= y_{k+1} + \gamma_{k+1}(y_{k+1} - y_k)
\end{align*}
\] with \( \gamma_{k+1} = \frac{t_k - 1}{t_{k+1}} \); \( t_k = \frac{k+a+1}{a} \) or \( \frac{1+\sqrt{1+4t_k^2}}{2} \).

with \( \alpha = \frac{1}{L} \),

\[
    t_{k+1}^2 F(y_{k+2}) - t_k^2 F(y_{k+1})
    \leq -\frac{1}{2\gamma} \| t_{k+1} y_{k+2} - (t_{k+1} - 1)y_{k+1} - y^* \|^2
    + \frac{1}{2\gamma} \| t_{k+1} x_{k+1} - (t_{k+1} - 1)y_{k+1} - y^* \|^2
\]

\[
    t_k^2 F(y_{k+1}) - t_{k-1}^2 F(y_k)
    \leq -\frac{1}{2\gamma} \| t_k y_{k+1} - (t_k - 1)y_k - y^* \|^2
    + \frac{1}{2\gamma} \| t_k x_k - (t_k - 1)y_k - y^* \|^2
\]

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:

\[
\begin{align*}
& x_k = y_k \\
& y_{k+1} = T_\alpha(x_k)
\end{align*}
\]

\[
\begin{align*}
x_{k+1} &= \text{extrapolation} \left( \{y_\ell \}_{\ell \leq k+1} \right) \\
y_{k+2} &= T_\alpha(x_{k+1})
\end{align*}
\]

**Choice 1**: \(1/k^2\) rate

\[
x_{k+1} = y_{k+1} - \frac{1}{t_{k+1}} (y_{k+1} - y_k) + \frac{t_{k-1}}{t_{k+1}} (y_k - y_{k-1})
\]

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

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

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

**Choice 2**: alternated inertia

\[
x_{k+1} = y_{k+1} + \gamma_{k+1} (y_{k+1} - y_k)
\]

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

- \(F(y_{2k})\) is non-increasing for \(\alpha = 1/L\) and \(\gamma_k \in [0, 1]\)
- Rate???
\( F \) is a KL function with \((F(u) - F^*)^{1-\theta} \leq C \cdot \text{dist}(0, \partial F(u))\) for all \( u : F(u) < F^* + \eta \) some \( C, \eta > 0, \theta \in (0, 1] \)

\( \mathcal{M} \) produce \((x_k)\) such that

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

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

<table>
<thead>
<tr>
<th>( \theta = 1 )</th>
<th>( a_k \geq a &gt; 0 )</th>
<th>( \gamma_k \leq \gamma &lt; 1 ) or stepsize ( &lt; 1/L )</th>
<th>finite number of steps</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta \in [0.5, 1[ )</td>
<td>( a_k = \frac{1}{k} )</td>
<td>Nesterov, ( d = 1 )</td>
<td>( \mathcal{O}\left(\left[\frac{C^2}{C^2+1}\right]^k\right))</td>
</tr>
<tr>
<td>( a_k = \frac{1}{k^d}, d \in ]0, 1[ )</td>
<td>( d \in ]0, 1[ )</td>
<td></td>
<td>( \mathcal{O}\left(\frac{1}{k}2C^2\right))</td>
</tr>
<tr>
<td>( \theta \in ]0, 0.5[ )</td>
<td>( a_k \geq a &gt; 0 )</td>
<td>( \gamma_k \leq \gamma &lt; 1 ) or stepsize ( &lt; 1/L )</td>
<td>( \mathcal{O}\left(\frac{1}{k^{1+\frac{2\theta}{1-2\theta}}}\right))</td>
</tr>
<tr>
<td>( a_k = \frac{1}{k} )</td>
<td>Nesterov, ( d = 1 )</td>
<td></td>
<td>( \mathcal{O}\left(\frac{1}{\log(k)1+\frac{2\theta}{1-2\theta}}\right))</td>
</tr>
<tr>
<td>( a_k = \frac{1}{k^d}, d \in ]0, 1[ )</td>
<td>( d \in ]0, 1[ )</td>
<td></td>
<td>( \mathcal{O}\left(\frac{1}{k^{1+\frac{2\theta-1+d}{1-2\theta}}}\right))</td>
</tr>
</tbody>
</table>
\( \ell_1 \) regularized logistic regression. \textit{ionosphere} dataset \((351 \times 35)\) 50% sparsity

![Graph showing functional error vs. number of iterations for different algorithms.](image)

- **Proximal gradient alg.**
- **FISTA**
- **MFISTA**

- **Alt. Iner. mono. \( d = 0.8 \)**
- **Alt. Iner. \( 1/k^2 \)**

1 \(/L_{\text{upper bound}}\) pessimistic stepsize
$\ell_1$ regularized logistic regression. Ionosphere dataset ($351 \times 35$) 50% sparsity

\begin{center}
\begin{tabular}{l}
Proximal gradient alg. \\
FISTA \\
MFISTA \\
Alt. Iner. mono. $d = 0.8$ \\
Alt. Iner. $1/k^2$
\end{tabular}
\end{center}

$\alpha = 8$ times less than the maximal stepsize for PG
\( \ell_1 \) regularized logistic regression. ionosphere dataset \((351 \times 35)\) 50% sparsity

\[
\begin{align*}
&\text{Proximal gradient alg.} \quad \text{FISTA} \quad \text{MFISTA} \\
&\text{Alt. Iner. mono. } d = 0.8 \quad \text{Alt. Iner. } 1/k^2
\end{align*}
\]

\( \alpha = 3 \) times less than the maximal stepsize for PG
$\ell_1$ regularized logistic regression. Ionosphere dataset ($351 \times 35$) 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