Relaxation and Inertia in Fixed-point iterations with Applications

Franck IUTZELER

LJK – Université Joseph Fourier

TITAN Workshop – November 27, 2015

Fixed Point Problem

Fixed Point Iterations

$$x^{k+1} = \mathsf{T}(x^k)$$

where T is some operator.

Sought Fixed Points

Fixed points of T:

$$x^\star = \mathsf{T}(x^\star)$$

but often, we actually want a *zero*:

$$\mathsf{T} = \mathsf{I} - \mathsf{U}; \mathsf{T} = (\mathsf{I} + \mathsf{U})^{-1}$$

Applications:

- Optimization: $\min_x f(x)$ with f convex and differentiable

Gradient algorithm $\rightarrow$ $\mathsf{T} = \mathsf{I} - \gamma\nabla f$.

Acceleration

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} \end{cases}$$

- Can we do better as we have:
  - Latest operation output: x^{k+1}
  - Past operation outputs: y^k, ...
  - Past iterates: x^k, ...
- Intuition: using current momentum to go further
- We will focus here on
  – *unit memory*
  – accelerations using past operation outputs OR iterates

We will name these mechanisms:

- relaxation $\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + (\eta - 1)(y^{k+1} - x^k) \end{cases}$
- inertia $\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + \gamma(y^{k+1} - y^k) \end{cases}$

Richardson Iterations (1910)

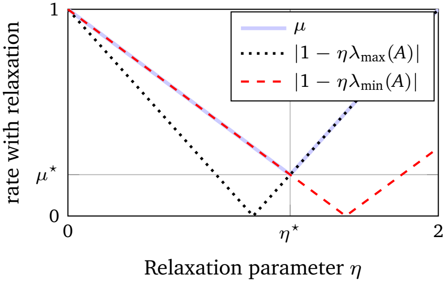
- Problem: Solve linear system $Ax = b$
- Iterations:

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

- Consequences: Faster linear (exponential) convergence rate for chosen η
- Related: Optimal

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

gives Chebyshev iterations.



Nesterov's Fast Gradient (1983)

- Problem: Solve minimization $\min_x f(x)$ with L-smooth f.
- Iterations:

$$\begin{cases} y^{k+1} = x^k - \frac{1}{L}\nabla f(x^k) \\ x^{k+1} = y^{k+1} + \gamma^{k+1}(y^{k+1} - y^k) \end{cases}$$

with $\gamma^{k+1} = \frac{\alpha^k - 1}{\alpha^{k+1}} \to 1$

where $\alpha^0 = 0$ and $\alpha^{k+1} = \frac{1+\sqrt{1+4(\alpha^k)^2}}{2}$.

- Consequences: Faster (sub-linear) convergence rate: $\mathcal{O}(1/k) \to \mathcal{O}(1/k^2)$
- Related: If f is in addition μ-strongly convex, fixed $\gamma = \frac{1-\sqrt{\mu/L}}{1+\sqrt{\mu/L}}$ is optimal to accelerate the linear convergence rate

FISTA (2008)

- Problem: Solve minimization $\min_x f(x) + g(x)$ with L-smooth f.
- Iterations:

$$\begin{cases} y^{k+1} = \arg\min_x \left\{ g(x) + \frac{L}{2} \left\| x - (x^k - \frac{1}{L}\nabla f(x^k)) \right\|^2 \right\} \\ x^{k+1} = y^{k+1} + \gamma^{k+1}(y^{k+1} - y^k) \end{cases}$$

with $\gamma^{k+1} = \frac{\alpha^k - 1}{\alpha^{k+1}} \to 1$
where $\alpha^0 = 0$ and $\alpha^{k+1} = \frac{1+\sqrt{1+4(\alpha^k)^2}}{2}$.

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

Fast *everything* (now)

Relaxation

- Equivalent to choosing stepsize in gradient algorithms
- Used in practice for ADMM but blurry guidelines "$\eta \in [1.5, 1.8]$ *usually speeds up the convergence*" [Eckstein1992]

Inertia

- Different variations studied [Tseng2008]
- Convergence of the *iterates* of FISTA is heavily investigated [Attouch2015]
- Different sequences used $\gamma^{k+1} = \frac{k}{k+\alpha}$, $\alpha > 3$ [Dossal2013]
- Same sequences used for ADMM but convergence is not guaranteed $\rightarrow$ restart mechanisms [O'Donohogue2013]

This talk

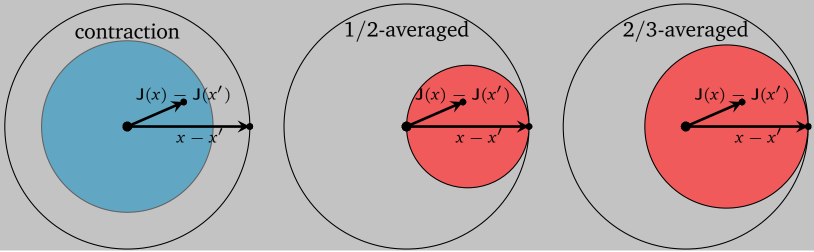
- Show relations and differences between relaxation and inertia
- Give different interpretations
- Derive some heuristics
- Numerical illustrations

Fixed point of Averaged operators

Fundamental Property

Let $\alpha \in]0,1[$. T is said to be *α-averaged* if

$$\forall x, x', \quad \|\mathsf{T}(x) - \mathsf{T}(x')\|^2 + \frac{1-\alpha}{\alpha}\|(\mathsf{I}-\mathsf{T})(x) - (\mathsf{I}-\mathsf{T})(x')\|^2 \leq \|x - x'\|^2$$



Examples:

- Gradient step: $\mathsf{T} = I - \frac{1}{L}\nabla f$ is $\frac{1}{2}$-averaged (FNE).
- Matrix multiplication: $\mathsf{T} = A$ is $\frac{1}{2}$-averaged if the eigenvalues of A are in the discus of center $1/2$ and radius $1/2$.

Convergence of Averaged operators

Fundamental Result: Krasnoselskiĭ-Mann iterations

Let $\alpha \in]0,1[$. If T is α-averaged, the sequence

$$x^{k+1} = \mathsf{T}(x^k)$$

converges to a fixed point of T.

The proof is based on the fact that x^k is *Fejér monotone* with respect to the fixed point of T:

$$\forall k, x^\star \in \mathbf{fix}\, \mathsf{T}, \qquad \|x^{k+1} - x^\star\| \leq \|x^k - x^\star\|$$

Relaxation and Inertia

Consider for k odd,

$$\begin{cases} x^{k+1} = \mathsf{T}_1(x^k + \nu^k(x^k - x^{k-1})) \\ x^{k+2} = \mathsf{T}_2(x^{k+1} + \nu^{k+1}(x^{k+1} - x^k)) \end{cases}$$

Relaxation

$\mathsf{T}_1 = \mathsf{T}$, $\mathsf{T}_2 = \mathsf{I}$

$\nu^k = 0$, $\nu^{k+1} = \eta^{k/2} - 1$

$x^{k+1} = \mathbf{T}(x^k)$

$x^{k+2} = x^{k+1} + (\eta^{k/2} - 1)(x^{k+1} - x^k)$

Inertia

$\mathsf{T}_1 = \mathsf{T}$, $\mathsf{T}_2 = \mathsf{T}$

$\nu^k = \gamma^k$, $\nu^{k+1} = \gamma^{k+1}$

$x^{k+1} = \mathbf{T}(x^k + \gamma^k(x^k - x^{k-1}))$

$x^{k+2} = \mathbf{T}(x^{k+1} + \gamma^{k+1}(x^{k+1} - x^k))$

Alternated Inertia

$\mathsf{T}_1 = \mathsf{T}$, $\mathsf{T}_2 = \mathsf{T}$

$\nu^k = 0$, $\nu^{k+1} = \gamma^{k+1}$

$x^{k+1} = \mathbf{T}(x^k)$

$x^{k+2} = \mathbf{T}(x^{k+1} + \gamma^{k+1}(x^{k+1} - x^k))$

Convergence of the iterates: relaxation

Convergence w/ relaxation

Let $\alpha \in]0,1[$. If T is α-averaged, the sequence

$$x^{k+1} = x^k + \eta^k(\mathsf{T}(x^k) - x^k)$$

converges to a fixed point of T under the condition that

$$0 < \underline{\eta} \leq \eta^k \leq \overline{\eta} < 1/\alpha.$$

- <u>Fejér monotone:</u> yes
- <u>Limit case:</u> $T([x,y]) = [x,0]$. Take $\eta = 1/\alpha = 2$, then

$$T_\eta([x,y]) = [2x - x; 0 - y] = [x, -y]$$

Convergence of the iterates: inertia

Convergence w/ inertia

Let $\alpha \in]0,1[$. If T is α-averaged, the sequence

$$x^{k+1} = \mathsf{T}(x^k + \gamma^k(x^k - x^{k-1}))$$

converges to a fixed point of T under the condition that either

$$i) \quad \exists \gamma, 0 \leq \gamma^k \leq \bar{\gamma} < 1 \quad \text{and} \quad \sum_{k=1}^{\infty} \gamma^k \|x^k - x^{k-1}\|^2 < \infty;$$

$$ii) \quad \gamma^k = \gamma \quad \text{and} \quad (1-\gamma)^2 > \frac{\alpha}{1-\alpha}\gamma(1+\gamma).$$

- Fejér monotone: NO. In general, inertia suffers from *not being monotonous*.
- Limit case: Take a fixed γ, if $\alpha = 1/2$, the condition for convergence becomes $\gamma < 1/3$.
  Consider the $1/2$-averaged operator equal to 0.5 identity plus a 0.5 times a small rotation $T = 0.5I + 0.5 \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$.

Convergence of the iterates: alternated inertia

Convergence w/ alternated inertia

Let $\alpha \in]0,1[$. If T is α-averaged, the sequence

$$\begin{cases} \mathsf{T}(x^k) & \text{if } k \text{ is even} \\ \mathsf{T}(x^k + \gamma^k(x^k - x^{k-1})) & \text{if } k \text{ is odd} \end{cases}$$

converges to a fixed point of T under the condition that

$$0 \leq \gamma^k \leq \frac{1-\alpha}{\alpha}.$$

- Fejér monotone: YES. Alternating inertia enables to retrieve monotonicity.
- Limit case: ???

Finite difference schemes

Consider the differential equality

$$\dot{x}(t) = -\nabla f(x(t))$$

with f smooth and convex.

- Explicit/Euler scheme:

$$\frac{x^{k+1} - x^k}{\gamma} = -\nabla f(x^k) \Rightarrow x^{k+1} = x^k - \gamma \nabla f(x^k)$$

- Implicit/Proximal scheme:

$$\begin{aligned} \frac{x^{k+1} - x^k}{\gamma} = -\nabla f(x^{k+1}) \Rightarrow x^{k+1} &= x^k - \gamma \nabla f(x^{k+1}) \\ &= \mathbf{prox}_{\gamma g}(x^k) \\ &= \arg\min_x f(x) + \frac{1}{2\gamma} \|x - x^k\|^2 \end{aligned}$$

- Implicit/Explicit schemes: $\dot{x}(t) = -\nabla f(x(t)) - \partial g(x(t))$
Explicit on f and implicit on g gives a forward backward algorithm.

Finite difference schemes II: 4th order Runge-Kutta

Consider the differential equality

$$\dot{x}(t) = -\nabla f(x(t))$$

with f smooth and convex.

- Runge-Kutta

$$\begin{aligned} g_1^{k+1} &= \nabla f(x^k) \\ g_2^{k+1} &= \nabla f(x^k - \frac{\gamma}{2} g_1^{k+1}) \\ g_3^{k+1} &= \nabla f(x^k - \frac{\gamma}{2} g_2^{k+1}) \\ g_4^{k+1} &= \nabla f(x^k - \gamma g_3^{k+1}) \end{aligned}$$

$$\bullet\ \frac{x^{k+1} - x^k}{\gamma} = -\frac{1}{6}\left(g_1^{k+1} + 2g_2^{k+1} + 2g_3^{k+1} + g_4^{k+1}\right)$$

$$\Rightarrow x^{k+1} = x^k - \frac{\gamma}{6}\left(g_1^{k+1} + 2g_2^{k+1} + 2g_3^{k+1} + g_4^{k+1}\right)$$

Never proximal/implicit and compatible with another (e.g. implicit) subgradient
Not beneficial on *simple* functions

Inertia as a second order term

Consider the differential equality

$$\ddot{x}(t) + \alpha(t)\dot{x}(t) = -\nabla f(x(t))$$

- Explicit/Implicit scheme:

$$\frac{x^{k+2} - 2x^{k+1} + x^k}{\gamma^2} + \alpha^k \frac{x^{k+1} - x^k}{\gamma} = -\nabla f(y^{k+1})$$

where y^{k+1} is a linear combination between x^k and x^{k+1} to be defined.

Inertia as a second order term

Consider the differential equality

$$\ddot{x}(t) + \alpha(t)\dot{x}(t) = -\nabla f(x(t))$$

- Explicit/Implicit scheme:

$$\frac{x^{k+2} - 2x^{k+1} + x^k}{\gamma^2} + \alpha^k \frac{x^{k+1} - x^k}{\gamma} = -\nabla f(y^{k+1})$$

where y^{k+1} is a linear combination between x^k and x^{k+1} to be defined.

$$x^{k+2} = \underbrace{x^{k+1} + (1 - \gamma\alpha^k)(x^{k+1} - x^k)} - \gamma^2 \nabla f(y^{k+1})$$

Inertia as a second order term

Consider the differential equality

$$\ddot{x}(t) + \alpha(t)\dot{x}(t) = -\nabla f(x(t))$$

- Explicit/Implicit scheme:

$$\frac{x^{k+2} - 2x^{k+1} + x^k}{\gamma^2} + \alpha^k \frac{x^{k+1} - x^k}{\gamma} = -\nabla f(y^{k+1})$$

where y^{k+1} is a linear combination between x^k and x^{k+1} to be defined.

$$x^{k+2} = \underbrace{x^{k+1} + (1 - \gamma\alpha^k)(x^{k+1} - x^k)}_{y^{k+1}} - \gamma^2 \nabla f(y^{k+1})$$

which leads to $\begin{cases} y^{k+1} = x^{k+1} + (1 - \gamma\alpha^k)(x^{k+1} - x^k) \\ x^{k+2} = y^{k+1} - \gamma^2 \nabla f(y^{k+1}) \end{cases}$

- $\alpha(t) = \alpha \rightarrow$ fixed inertia; $\alpha(t) = \alpha/t \rightarrow \quad \gamma^k = \frac{k-1}{k+\alpha-1}$.
- Used very recently [Attouch2015] to prove iterates convergence of accelerated Forward-Backward

Averaged linear operators

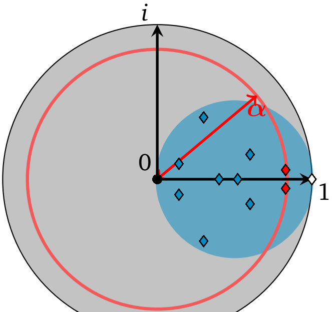
Theorem

Let $\alpha \in]0,1[$. Suppose that operator $\mathsf{T} = R \cdot + d$ is α-averaged and has fixed points. Then,

- the eigenvalues of R are in the disk of center $1-\alpha$ and radius α
- the sequence $(x^k)_{k>0}$ generated by $x^{k+1} = \mathsf{T}(x^k)$ converges linearly to a fixed point of T at rate

$$\gamma \triangleq \max\{|\lambda| : \lambda \neq 1 \text{ is an eigenvalue of } R\}$$

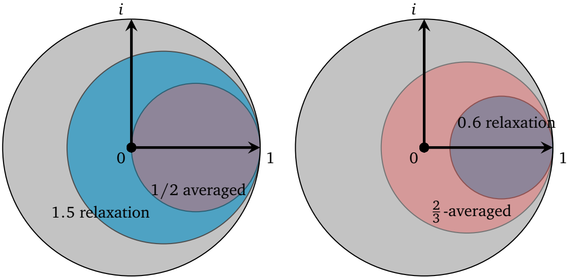
Example for a 1/2-averaged operator:



Effects of Relaxation and Inertia on eigenvalues

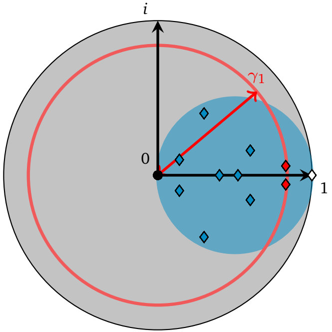
- Modifies the eigenvalues and thus rate
- Intuition possible for relaxation, less for inertia

Effect of relaxation on eigenvalues:
- inflation
- translation



Effects of Relaxation on eigenvalues

Back to our example of a 1/2-averaged operator:



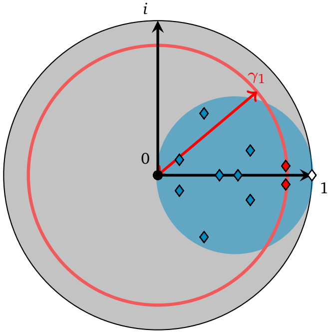
How to choose the relaxation parameter η ?

$\eta < 1$
under-relaxation

$\eta > 1$
over-relaxation

Effects of Relaxation on eigenvalues

Back to our example of a $1/2$-averaged operator:



How to choose the relaxation parameter η ?

$\eta < 1$
under-relaxation

$\eta > 1$
over-relaxation

Over-relaxation is in general beneficial for $1/2$ averaged operators

Another Simplification

$\mathsf{T} = R \cdot +d$ with R symmetric

- R has *real eigenvalues*
- T is *cyclically monotone*
- T is a *gradient step* on a convex function

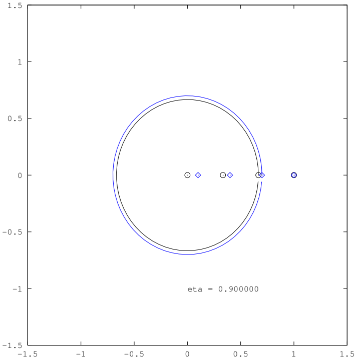
These points are equivalent from [Chiu1976] and Baillon-Haddad theorem

What are the optimal relaxation and inertia parameters ?

- Compute rate versus parameter $\mu(\cdot)$
- Find optimal parameter $\cdot^{\star}$ and rate $\mu^{\star}$

Relaxation

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + (\eta - 1)(y^{k+1} - x^k) \end{cases}$$

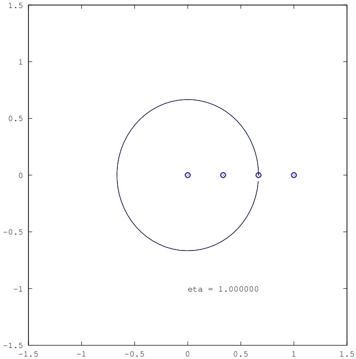


$$\eta^\star = \frac{2}{2\alpha + 1 - \lambda}$$
$$\mu^\star = \frac{2\alpha - 1 + \lambda}{2\alpha + 1 - \lambda}$$

- Depends on *extremal* eigenvalues

Relaxation

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + (\eta - 1)(y^{k+1} - x^k) \end{cases}$$

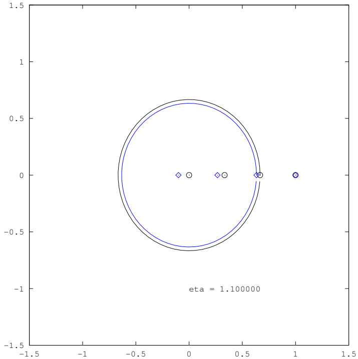


$$\eta^\star = \frac{2}{2\alpha + 1 - \lambda}$$
$$\mu^\star = \frac{2\alpha - 1 + \lambda}{2\alpha + 1 - \lambda}$$

- Depends on *extremal* eigenvalues

Relaxation

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + (\eta - 1)(y^{k+1} - x^k) \end{cases}$$



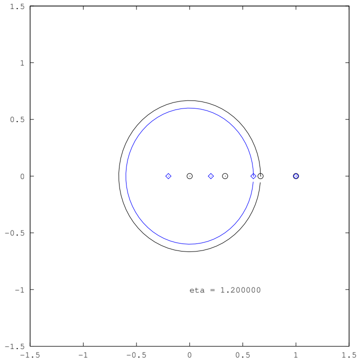
$$\eta^\star = \frac{2}{2\alpha + 1 - \lambda}$$

$$\mu^\star = \frac{2\alpha - 1 + \lambda}{2\alpha + 1 - \lambda}$$

- Depends on *extremal* eigenvalues

Relaxation

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + (\eta - 1)(y^{k+1} - x^k) \end{cases}$$



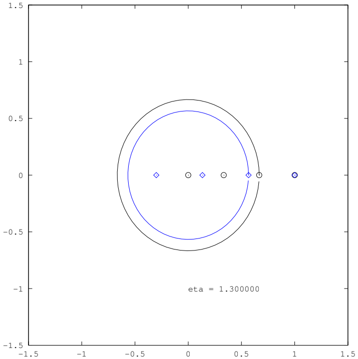
$$\eta^\star = \frac{2}{2\alpha + 1 - \lambda}$$

$$\mu^\star = \frac{2\alpha - 1 + \lambda}{2\alpha + 1 - \lambda}$$

- Depends on *extremal* eigenvalues

Relaxation

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + (\eta - 1)(y^{k+1} - x^k) \end{cases}$$



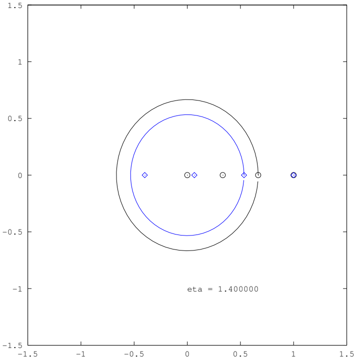
$$\eta^\star = \frac{2}{2\alpha + 1 - \lambda}$$

$$\mu^\star = \frac{2\alpha - 1 + \lambda}{2\alpha + 1 - \lambda}$$

- Depends on *extremal* eigenvalues

Relaxation

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + (\eta - 1)(y^{k+1} - x^k) \end{cases}$$



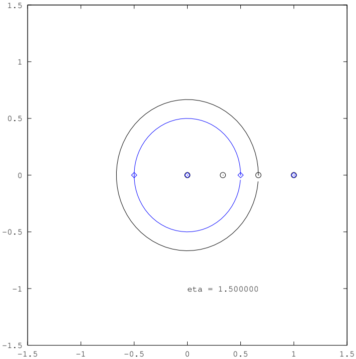
$$\eta^\star = \frac{2}{2\alpha + 1 - \lambda}$$

$$\mu^\star = \frac{2\alpha - 1 + \lambda}{2\alpha + 1 - \lambda}$$

- Depends on *extremal* eigenvalues

Relaxation

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + (\eta - 1)(y^{k+1} - x^k) \end{cases}$$



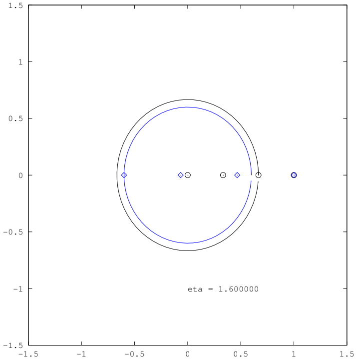
$$\eta^\star = \frac{2}{2\alpha + 1 - \lambda}$$

$$\mu^\star = \frac{2\alpha - 1 + \lambda}{2\alpha + 1 - \lambda}$$

- Depends on *extremal* eigenvalues

Relaxation

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + (\eta - 1)(y^{k+1} - x^k) \end{cases}$$



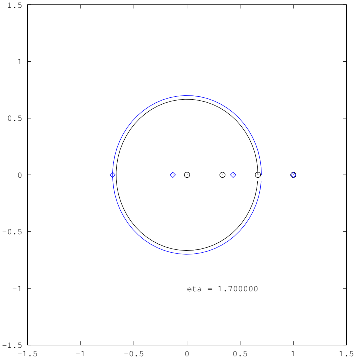
$$\eta^\star = \frac{2}{2\alpha + 1 - \lambda}$$

$$\mu^\star = \frac{2\alpha - 1 + \lambda}{2\alpha + 1 - \lambda}$$

- Depends on *extremal* eigenvalues

Relaxation

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + (\eta - 1)(y^{k+1} - x^k) \end{cases}$$



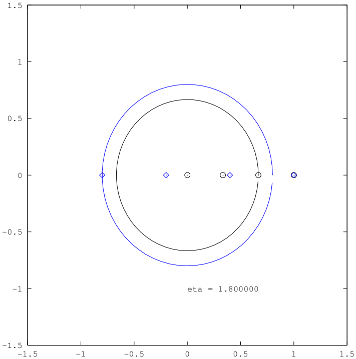
$$\eta^\star = \frac{2}{2\alpha + 1 - \lambda}$$

$$\mu^\star = \frac{2\alpha - 1 + \lambda}{2\alpha + 1 - \lambda}$$

- Depends on *extremal* eigenvalues

Relaxation

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + (\eta - 1)(y^{k+1} - x^k) \end{cases}$$

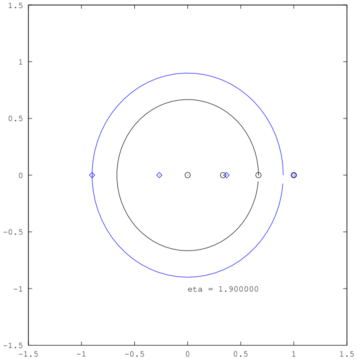


$$\eta^\star = \frac{2}{2\alpha + 1 - \lambda}$$
$$\mu^\star = \frac{2\alpha - 1 + \lambda}{2\alpha + 1 - \lambda}$$

- Depends on *extremal* eigenvalues

Relaxation

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + (\eta - 1)(y^{k+1} - x^k) \end{cases}$$



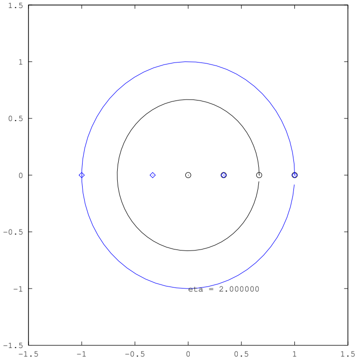
$$\eta^\star = \frac{2}{2\alpha + 1 - \lambda}$$

$$\mu^\star = \frac{2\alpha - 1 + \lambda}{2\alpha + 1 - \lambda}$$

- Depends on *extremal* eigenvalues

Relaxation

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + (\eta - 1)(y^{k+1} - x^k) \end{cases}$$



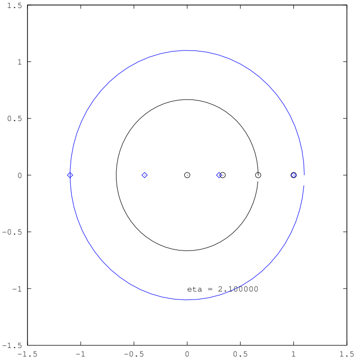
$$\eta^\star = \frac{2}{2\alpha + 1 - \lambda}$$

$$\mu^\star = \frac{2\alpha - 1 + \lambda}{2\alpha + 1 - \lambda}$$

- Depends on *extremal* eigenvalues

Relaxation

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + (\eta - 1)(y^{k+1} - x^k) \end{cases}$$



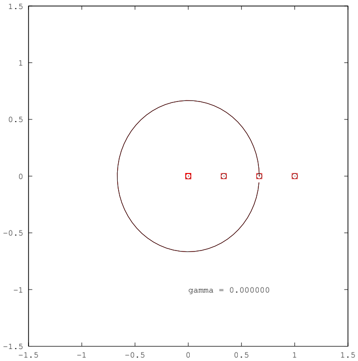
$$\eta^\star = \frac{2}{2\alpha + 1 - \lambda}$$

$$\mu^\star = \frac{2\alpha - 1 + \lambda}{2\alpha + 1 - \lambda}$$

- Depends on *extremal* eigenvalues

Inertia

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + \gamma(y^{k+1} - y^k) \end{cases}$$



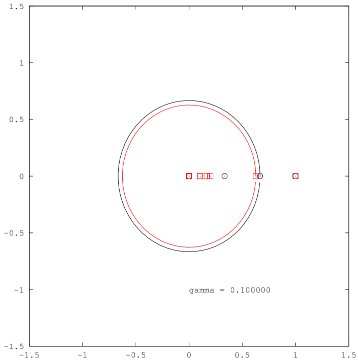
for $\alpha \leq 1/2$

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

- Depends on the *largest* eigenvalue
- Penalized by negative eigenvalues

Inertia

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + \gamma(y^{k+1} - y^k) \end{cases}$$



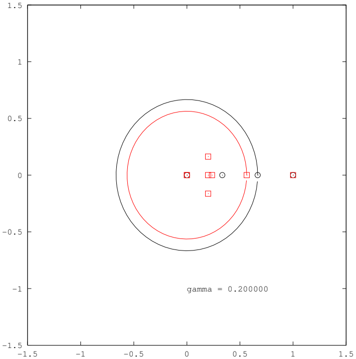
for $\alpha \leq 1/2$

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

- Depends on the *largest* eigenvalue
- Penalized by negative eigenvalues

Inertia

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + \gamma(y^{k+1} - y^k) \end{cases}$$



for $\alpha \leq 1/2$

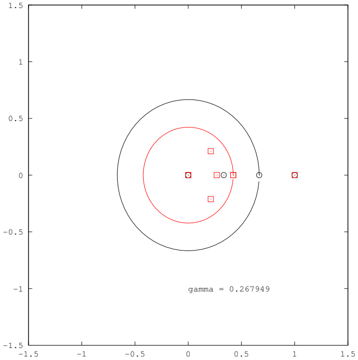
$$\gamma^\star = \frac{(1-\sqrt{1-\lambda})^2}{\lambda}$$

$$\mu^\star = 1 - \sqrt{1-\lambda}$$

- Depends on the *largest* eigenvalue
- Penalized by negative eigenvalues

Inertia

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + \gamma(y^{k+1} - y^k) \end{cases}$$



for $\alpha \leq 1/2$

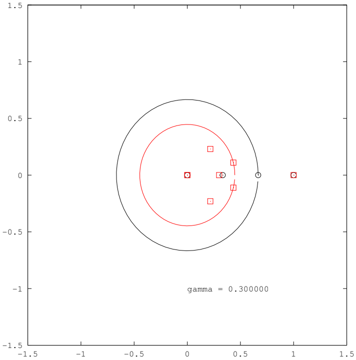
$$\gamma^\star = \frac{(1 - \sqrt{1-\lambda})^2}{\lambda}$$

$$\mu^\star = 1 - \sqrt{1-\lambda}$$

- Depends on the *largest* eigenvalue
- Penalized by negative eigenvalues

Inertia

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + \gamma(y^{k+1} - y^k) \end{cases}$$



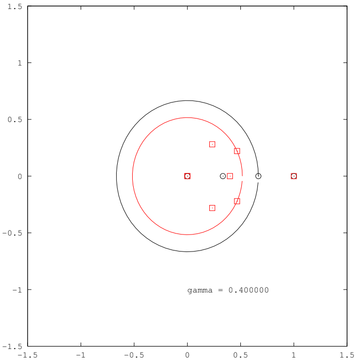
for $\alpha \leq 1/2$

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

- Depends on the *largest* eigenvalue
- Penalized by negative eigenvalues

Inertia

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + \gamma(y^{k+1} - y^k) \end{cases}$$



for $\alpha \leq 1/2$

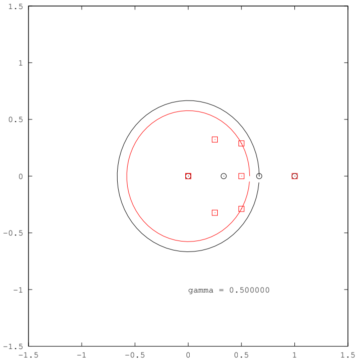
$$\gamma^\star = \frac{(1-\sqrt{1-\lambda})^2}{\lambda}$$

$$\mu^\star = 1 - \sqrt{1-\lambda}$$

- Depends on the *largest* eigenvalue
- Penalized by negative eigenvalues

Inertia

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + \gamma(y^{k+1} - y^k) \end{cases}$$



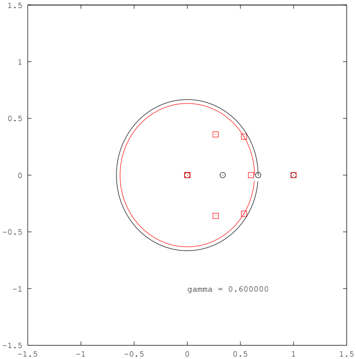
for $\alpha \leq 1/2$

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

- Depends on the *largest* eigenvalue
- Penalized by negative eigenvalues

Inertia

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + \gamma(y^{k+1} - y^k) \end{cases}$$



for $\alpha \leq 1/2$

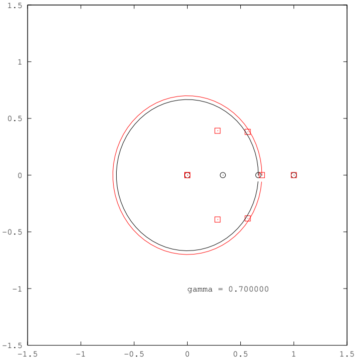
$$\gamma^\star = \frac{(1 - \sqrt{1-\lambda})^2}{\lambda}$$

$$\mu^\star = 1 - \sqrt{1-\lambda}$$

- Depends on the *largest* eigenvalue
- Penalized by negative eigenvalues

Inertia

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + \gamma(y^{k+1} - y^k) \end{cases}$$



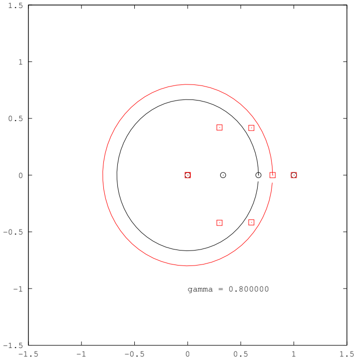
for $\alpha \leq 1/2$

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

- Depends on the *largest* eigenvalue
- Penalized by negative eigenvalues

Inertia

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + \gamma(y^{k+1} - y^k) \end{cases}$$



for $\alpha \leq 1/2$

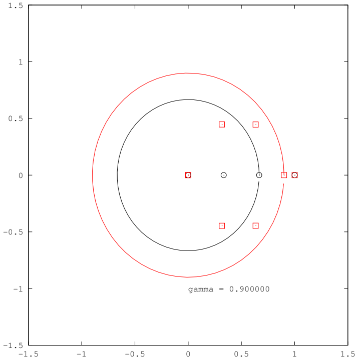
$$\gamma^\star = \frac{(1 - \sqrt{1-\lambda})^2}{\lambda}$$

$$\mu^\star = 1 - \sqrt{1-\lambda}$$

- Depends on the *largest* eigenvalue
- Penalized by negative eigenvalues

Inertia

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + \gamma(y^{k+1} - y^k) \end{cases}$$



for $\alpha \leq 1/2$

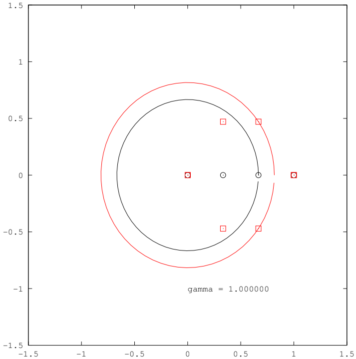
$$\gamma^\star = \frac{(1 - \sqrt{1-\lambda})^2}{\lambda}$$

$$\mu^\star = 1 - \sqrt{1-\lambda}$$

- Depends on the *largest* eigenvalue
- Penalized by negative eigenvalues

Inertia

$$\begin{cases} y^{k+1} = \mathsf{T}(x^k) \\ x^{k+1} = y^{k+1} + \gamma(y^{k+1} - y^k) \end{cases}$$



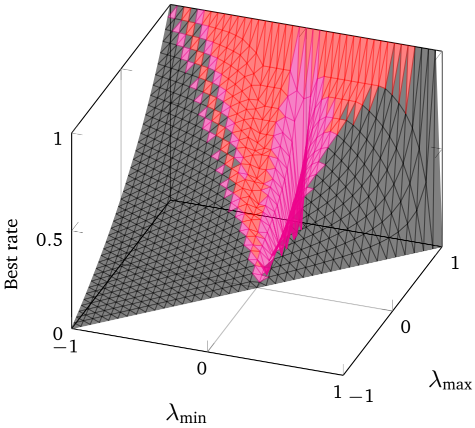
for $\alpha \leq 1/2$

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

$$\mu^\star = 1 - \sqrt{1-\lambda}$$

- Depends on the *largest* eigenvalue
- Penalized by negative eigenvalues

Comparison of relaxation and (alternated) inertia



- Two eigenvalues $\lambda_{\min} < \lambda_{\max}$
- Best rate by either *relaxation*, *inertia*, or *alternated inertia*
- *alternated inertia*: best when $\lambda_{\min} \approx 0$
- *inertia*: best when $\lambda_{\max} \approx 1$

Conclusions

- 3 different behaviors for the eigenvalues
- *(alternated) inertia* is penalized for negative eigenvalues i.e. $\alpha > 1/2$
- Finding the optimal parameters $\approx$ finding eigenvalues of R
  $\approx$ as hard as the original problem

Development of Online Acceleration Algorithms: Work in Progress: contact me!

i) Estimate current speed (linear rate)

ii) Retrieve original (unaccelerated) speed

iii) Take the tuned parameter equal to optimal one with $\lambda =$ original rate

Online Relaxation Method

Online Relaxation Method (ORM)
for an α-averaged operator T :
Initialization:
$\varepsilon \in]0, 2\min(\alpha; 1-\alpha)], x^0, x^1 = \mathsf{T}x^0 , \eta^0 = \eta^1 = 1.$
At each iteration $k \geq 1$:

$$\eta^{k+1} = \frac{(2-\varepsilon)\eta^k}{2\alpha\eta^k + 1 - \frac{\eta^{k-1}\|x^k - x^{k-1}\|}{\eta^k\|x^{k-1}-x^{k-2}\|}} + \frac{\varepsilon}{4\alpha}$$

$$x^{k+1} = \eta^{k+1}\mathsf{T}x^k + (1-\eta^{k+1})x^k$$

- Complexity equivalent to original algorithm
- Convergence ensured by the ε for any α-averaged operator

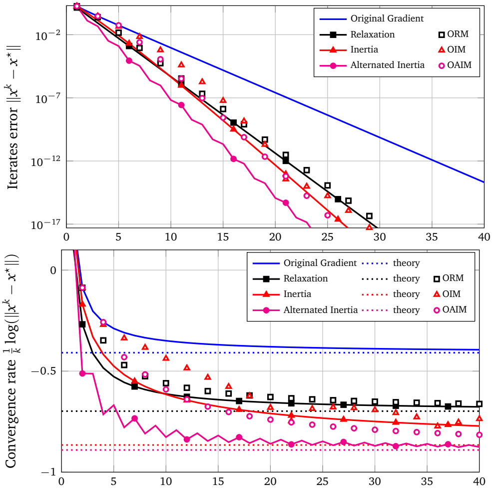
Gradient on a quadratic function

- Problem: $\min_x x^T Q x + b^T x$ with Q positive definite
  - Lipchitz constant L is *known*
  - Strong convexity constant $\mu > 0$ is *unknown*

Relaxation	*Inertia*	*Alternated Inertia*
$x^{k+1} = x^k - \frac{\eta^{k+1}}{L} \nabla F(x^k)$	$\begin{cases} y^k = x^k + \gamma^k (x^k - x^{k-1}) \\ x^{k+1} = y^k - \frac{1}{L} \nabla f(y^k) \end{cases}$	$\begin{cases} y^k = x^k + \gamma^k (x^k - x^{k-1}) \\ x^{k+1} = y^k - \frac{1}{L} \nabla f(y^k) \\ x^{k+2} = x^{k+1} - \frac{1}{L} \nabla f(x^{k+1}) \end{cases}$
$\eta^\star = \frac{2}{1+\mu/L}$	$\gamma^\star = \frac{1-\sqrt{\mu/L}}{1+\sqrt{\mu/L}}$	$\frac{2(\mu/L)^2 - (3+\sqrt{2})\mu/L + 1 + \sqrt{2}}{-2(\mu/L)^2 + 2\mu/L + \frac{1}{2}}$
$\mu^\star = \frac{1-\mu/L}{1+\mu/L}$	$\mu^\star = 1 - \sqrt{\mu/L}$	$\mu^\star = \frac{(\gamma^\star)^2}{4(1+\gamma^\star)}$

- *Relaxation*: (μ, L)-smooth function optimal stepsize [Taylor2015]
- *Inertia*: (μ, L)-smooth function optimal inertia [Nesterov2004]
- Accelerations are *not* additive (relaxation gives negative eig.)
- Best rates:
  Alternated Inertia for $\frac{4}{9+4\sqrt{2}} \approx 0.273 \leq \frac{\mu}{L} \leq 1$
  Inertia elsewhere
  but parameters outside contraction-based convergence results if $\frac{\mu}{L} \leq 1/4$

Gradient on a quadratic function



Alternating Direction Method of Multipliers (ADMM)

- <u>Problem:</u> $\min_{x\in\mathbb{R}^n} f(x) + g(Mx)$

Alternating Direction Method of Multipliers (ADMM)

$$x^{k+1} = \operatorname*{argmin}_w \left\{ f(w) + \frac{\rho}{2} \left\| Mw - z^k + \frac{\lambda^k}{\rho} \right\|^2 \right\}$$

$$z^{k+1} = \operatorname*{argmin}_w \left\{ g(w) + \frac{\rho}{2} \left\| Mx^{k+1} - w + \frac{\lambda^k}{\rho} \right\|^2 \right\}$$

$$\lambda^{k+1} = \lambda^k + \rho(Mx^{k+1} - z^{k+1})$$

- For an operator point of view, it is 1/2-averaged

Alternating Direction Method of Multipliers (ADMM)

- Problem: $\min_{x\in\mathbb{R}^n} f(x) + g(Mx)$

Relaxed ADMM

$$x^{k+1} = \operatorname*{argmin}_{w} \left\{ f(w) + \frac{\rho}{2}\left\| Mw - z^k + \frac{\lambda^k}{\rho} \right\|^2 \right\}$$

$$z^{k+1} = \operatorname*{argmin}_{w} \left\{ g(w) + \frac{\rho}{2}\left\| \eta M x^{k+1} + (1-\eta) z^k - w + \frac{\lambda^k}{\rho} \right\|^2 \right\}$$

$$\lambda^{k+1} = \lambda^k + \rho(\eta M x^{k+1} + (1-\eta) z^k - z^{k+1})$$

Alternating Direction Method of Multipliers (ADMM)

- Problem: $\min_{x\in\mathbb{R}^n} f(x) + g(Mx)$

Inertial ADMM

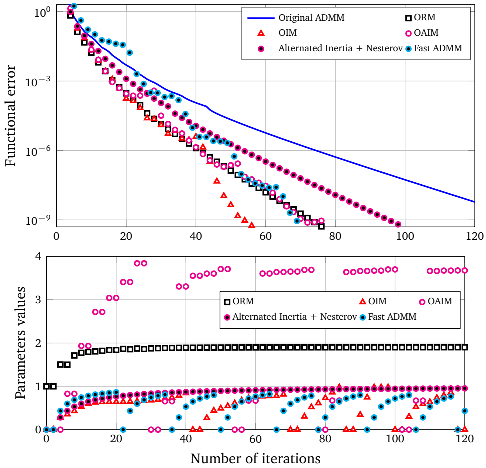
$$x^{k+1} = \operatorname*{argmin}_w \left\{ f(w) + \frac{\rho}{2}\left\| Mw - z^k + \frac{\lambda^k}{\rho} \right\|^2 \right\}$$

$$z^{k+1} = \operatorname*{argmin}_w \left\{ g(w) + \frac{\rho}{2}\left\| Mx^{k+1} - w + \frac{\lambda^k}{\rho} + \gamma\left(M(x^{k+1} - x^k) + \frac{\lambda^k - \lambda^{k-1}}{\rho} \right) \right\|^2 \right\}$$

$$\lambda^{k+1} = \lambda^k + \rho(Mx^{k+1} - z^{k+1}) + \gamma\rho\left(M(x^{k+1} - x^k) + \frac{\lambda^k - \lambda^{k-1}}{\rho} \right)$$

ADMM on LASSO problem

- Problem: $\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1$



Conclusion

- Different methods to accelerate *
- Analysis quickly becomes involved
- Alternated Inertia surprisingly performing
- Attractive Online Algorithms to accelerate for no complexity
- **Work in Progress**: contact me!

- Stability improvement for inertia
- Combinations of relaxation/inertia
- Acceleration on randomized algorithms
- Distributed algorithms

Merci de votre attention.

Franck IUTZELER – `www.iutzeler.org`