Optimization algorithms and differential equations: theory and insights

K.C Zygalakis

School of Mathematics, University of Edinburgh

Maxwell Institute for Mathematical Sciences

Workshop on Recent Advances in Iterative Reconstruction
University College London

Collaborators

Jesus M. Sanz-Serna (UC3M)

Paul Dobson (Edinburgh)

- J. M. Sanz-Serna and K. C. Zygalakis. The connections between Lyapunov functions for some optimization algorithms and differential equations. *SIAM Journal on Numerical Analysis*, 59(3), 1542–1565, 2021.
- P. Dobson, J. M. Sanz-Serna and K. C. Zygalakis, On the connections between optimization algorithms, Lyapunov functions, and differential equations: theory and insights, arXiv:2305.08658, (2023)

Engineering and Physical Sciences Research Council

The Leverhulme Trust

Overview

THE UNIVERSITY of EDINBURGH

Overview

THE UNIVERSITY *of* EDINBURGH

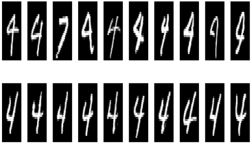
Statement of an innocent looking problem

Optimization

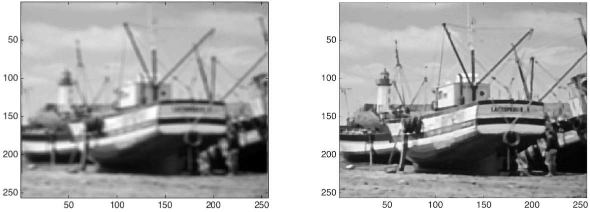
Find the unconstrained minimum of a function $\pi(x)$ in $\mathbb{R}^d$

$$\min_{x \in \mathbb{R}^d} \pi(x)$$

Numerous applications



(a) Image classification



(b) Image reconstruction

Gradient flow

Consider the differential equation:

$$\frac{dx}{dt} = -\nabla\pi(x).$$

This has the interesting property that

$$\frac{d\pi(x)}{dt} = -\left\|\nabla\pi(x)\right\|^2 \Rightarrow \lim_{t\to\infty} x(t) = x^*,$$

where x^* is a (unique) minimizer. This makes the equation above central (or at least the simplest choice) for optimization purposes.

In an ideal world!!!

- There is nothing to be done...
- Discretize the candidate differential equations and go
    - *Optimization*: Go to infinity as quickly as possible (in terms of function evaluations).

In real life...

- Starting from the differential equation and discretising might not be enough in terms of mimicking the rate of convergence to equilibrium.
- Going to infinity as quickly as possible implies that you can use arbitrary large time-steps in your numerical discretization.
- Reality unfortunately comes back to bite you, as time-steps restrictions appear once you discretize your differential equation.

Overview

Optimization: Continuous case

Gradient flow:

$$\dot{x} + \nabla f(x) = 0$$

Momentum equation:

$$\ddot{x} + \bar{b}\sqrt{m}\dot{x} + \nabla f(x) = 0$$

Quadratic case: $f(x) = \frac{1}{2}x^T Qx, \quad \sigma(Q) \in [m, L]$

Nonlinear case: $f(x) \in \mathcal{F}(m, L)$

[1] W. Su, S. Boyd, E. J. Candés NIPS 2014: 2510-2518, (2014).

Continuous time formulation

$$\dot{\xi}(t) = \bar{A}\xi(t) + \bar{B}u(t),$$
$$y(t) = \bar{C}\xi(t),$$
$$u(t) = \nabla f(y(t)).$$

where $\xi(t) \in \mathbb{R}^n$ is the state, $y(t) \in \mathbb{R}^d (d \leq n)$ the output, and $u(t) = \nabla f(y(t))$ the continuous feedback input. Fixed points of the system satisfy

$$0 = \bar{A}\xi^\star, \quad y^\star = \bar{C}\xi^\star, \quad u^\star = \nabla f(y^\star);$$

in our context $u^\star = 0$ and $y^\star = x^\star$.

[2] M.Fazlyab, A. Ribeiro, M. Morari, V. M. Preciado, *SIAM J. Optim.*, 28(3), 2654–2689, (2018).

Examples

1. Gradient flow: $\dot{x} = -\nabla f(x)$.

$$\bar{A} = 0_{d\times d}, \quad \bar{B} = -I_{d\times d}, \quad \bar{C} = I_{d\times d}.$$

2. Momentum equation: $\ddot{x} + \bar{b}\sqrt{m}\dot{x} + \nabla f(x) = 0$.

$$\bar{A} = \begin{bmatrix} -\bar{b}\sqrt{m}I_d & 0_d \\ \sqrt{m}I_d & 0_d \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} -(1/\sqrt{m})I_d \\ 0_d \end{bmatrix}, \quad \bar{C} = \begin{bmatrix} 0_d & I_d \end{bmatrix}.$$

THE UNIVERSITY of EDINBURGH

Quadratic case

- The continuous time formulation now becomes

$$\dot{\xi}(t) = (\bar{A} + \bar{B}Q\bar{C})\xi(t)$$

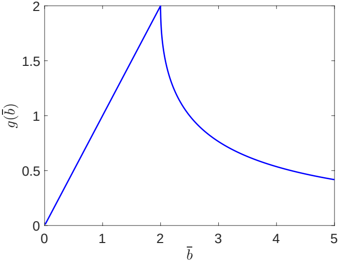
- Solution is given by

$$\xi(t) = e^{(\bar{A}+\bar{B}Q\bar{C})t}\xi(0)$$

- To deduce a convergence rate to the minimizer we need to understand the spectral properties of $e^{(\bar{A}+\bar{B}Q\bar{C})t}$

Quadratic case: Gradient flow vs momentum equations

- Gradient flow: rate of convergence e^{-2mt}
- Momentum equation: rate of convergence $e^{-g(\bar{b})\sqrt{m}t}$



- Clearly using the first order dynamics is suboptimal in terms of convergence

The class $\mathcal{F}(m, L)$

1. $\langle x - y, \nabla f(x) - \nabla f(y) \rangle \geq m \|x - y\|^2$.
2. $\|\nabla f(x) - \nabla f(y)\|^2 \leq L^2 \|x - y\|^2$.
3. $\frac{mL}{m+L} \|x - y\|^2 + \frac{1}{m+L} \|\nabla f(x) - \nabla f(y)\|^2 \leq (\nabla f(x) - \nabla f(y))^T (x - y)$

An equivalent way of expressing these equations are the following quadratic constraints:

1. $$\begin{bmatrix} x - y \\ \nabla f(x) - \nabla f(y) \end{bmatrix}^T \begin{bmatrix} -\frac{m}{2} I_d & \frac{1}{2} I_d \\ \frac{1}{2} I_d & 0_d \end{bmatrix} \begin{bmatrix} x - y \\ \nabla f(x) - \nabla f(y) \end{bmatrix} \geq 0.$$
2. $$\begin{bmatrix} x - y \\ \nabla f(x) - \nabla f(y) \end{bmatrix}^T \begin{bmatrix} L^2 I_d & 0_d \\ 0_d & -I_d \end{bmatrix} \begin{bmatrix} x - y \\ \nabla f(x) - \nabla f(y) \end{bmatrix} \geq 0.$$
3. $$\begin{bmatrix} x - y \\ \nabla f(x) - \nabla f(y) \end{bmatrix}^T \begin{bmatrix} -\frac{mL}{m+L} I_d & \frac{1}{2} I_d \\ \frac{1}{2} I_d & -\frac{1}{m+L} I_d \end{bmatrix} \begin{bmatrix} x - y \\ \nabla f(x) - \nabla f(y) \end{bmatrix} \geq 0.$$

(Continuous) Lyapunov functions

Consider

$$V(\xi(t), t) = \alpha(t)(f(y(t)) - f(y_*)) + (\xi(t) - \xi_*)P(t)(\xi(t) - \xi_*)$$

and assume that we can find $\alpha(t), P(t) \succeq 0$ such that

$$V(\xi(t), t) \leq V(\xi(t_0), t_0)$$

then

$$0 \leq f(y(t)) - f(y_*) \leq V(\xi(t_0, t_0))/\alpha(t) = \mathcal{O}(1/\alpha(t))$$

A small calculation

By differentiating the Lyapunov function we have

$$\begin{aligned}\dot{V} &= \dot{\alpha}(t)(f(y(t)) - f(y_*)) \\ &\quad + \alpha(t)(\nabla f(y(t)) - \nabla f(y_*))^T \dot{y}(t) \\ &\quad + 2(\xi(t) - \xi_*)^T P(t)\dot{\xi}(t) \\ &\quad + (\xi(t) - \xi_*)^T \dot{P}(t)(\xi(t) - \xi_*)^T\end{aligned}$$

Setting $e(t) = [(\xi(t) - \xi_*)^T (u(t) - u_*)^T]$ and using the strong convexity properties of f ($f \in \mathcal{F}_{m,L}$) we can obtain

$$\dot{V}(t) \leq e^T(t)(\cdots)e(t)$$

and if the matrix inside the parenthesis is negative definite then we are done.

THE UNIVERSITY of EDINBURGH

A theorem for the (continuous) Lyapunov function

(Continuous) convergence to the minimizer

Suppose that there exist $\lambda > 0$, $\bar{P} \succeq 0$, and $\sigma \geq 0$ that satisfy

$$\bar{T} = \bar{M}^{(0)} + \bar{M}^{(1)} + \lambda \bar{M}^{(2)} + \sigma \bar{M}^{(3)} \preceq 0$$

where

$$\bar{M}^{(0)} = \begin{bmatrix} \bar{P}\bar{A} + \bar{A}^{\mathcal{T}}\bar{P} + \lambda\bar{P} & \bar{P}\bar{B} \\ \bar{B}^{\mathcal{T}}\bar{P} & 0 \end{bmatrix},$$

$$\bar{M}^{(1)} = \frac{1}{2}\begin{bmatrix} 0 & (\bar{C}\bar{A})^{\mathcal{T}} \\ \bar{C}\bar{A} & \bar{C}\bar{B} + \bar{B}^{\mathcal{T}}\bar{C}^{\mathcal{T}} \end{bmatrix},$$

$$\bar{M}^{(2)} = \begin{bmatrix} \bar{C}^{\mathcal{T}} & 0 \\ 0 & I_d \end{bmatrix} \begin{bmatrix} -\frac{m}{2} I_d & \frac{1}{2} I_d \\ \frac{1}{2} I_d & 0 \end{bmatrix} \begin{bmatrix} \bar{C} & 0 \\ 0 & I_d \end{bmatrix},$$

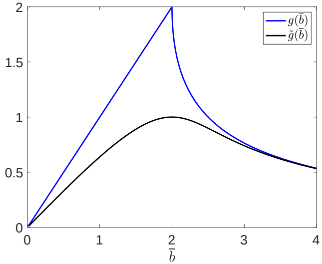
$$\bar{M}^{(3)} = \begin{bmatrix} \bar{C}^{\mathcal{T}} & 0 \\ 0 & I_d \end{bmatrix} \begin{bmatrix} -\frac{mL}{m+L} I_d & \frac{1}{2} I_d, \\ \frac{1}{2} I_d & -\frac{1}{m+L} I_d \end{bmatrix} \begin{bmatrix} \bar{C} & 0 \\ 0 & I_d \end{bmatrix}.$$

Then the following inequality holds for $f \in \mathcal{F}_{m,L}$, $t \geq 0$,

$$f(y(t)) - f(y^\star) \leq e^{-\lambda t}\left(f(y(0)) - f(y^\star) + (\xi(0) - \xi^\star)^{\mathcal{T}}\bar{P}(\xi(0) - \xi^\star)\right).$$

Nonlinear case: Gradient flow vs momentum equations

- Gradient flow: Again we have that $\lambda = 2m$.
- Momentum equations: We have that $\lambda = \tilde{g}(\bar{b})\sqrt{m}$



1. You lose some of the rate you can prove between the linear and the nonlinear case
2. Still the momentum dynamics accelerate the convergence to equilibrium ($\sqrt{m} \gg m$ when $m \ll 1$.)
3. One should discretise the momentum dynamics.

THE UNIVERSITY of EDINBURGH

Discrete time

$$
\begin{aligned}
\xi_{k+1} &= A\xi_k + Bu_k, \\
u_k &= \nabla f(y_k), \\
y_k &= C\xi_k, \\
x_k &= E\xi_k.
\end{aligned}
$$

A family of algorithms

$$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha \nabla f(y_k),$$
$$y_k = x_k + \gamma(x_k - x_{k-1}),$$

1. For $\beta = \gamma = 0$ we recover the gradient descent
$$x_{k+1} = x_k - \alpha \nabla f(x_k).$$
2. For $\gamma = \beta$ we recover the Nesterov method.
3. For $\gamma = 0$, $\beta \neq 0$ we recover the heavy ball method.

Quadratic case

- The continuous time formulation now becomes

$$\xi_{k+1} = (A + BQC)\xi_k$$

- Solution is given by

$$\xi_k = (A + BQC)^k \xi(0)$$

- To deduce a convergence rate to the minimizer we need to understand the spectral properties of $(A + BQC)$

Quadratic case: Convergence rates

$$\|\xi_k - \xi^*\|^2 \leq \rho^{2k} \|\xi_0 - \xi^*\|^2$$

1. Gradient descent: $\alpha = \frac{2}{m+L}$, and $\rho = \frac{\kappa-1}{\kappa+1}$
2. Nesterov method: $\alpha = \frac{4}{3L+m}$, $\beta = \frac{\sqrt{3\kappa+1}-2}{\sqrt{3\kappa+1}+2}$, and $\rho = 1 - \frac{2}{\sqrt{3\kappa+1}}$
3. Heavy ball: $\alpha = \frac{4}{(\sqrt{L}+\sqrt{m})^2}$, $\beta = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$, and $\rho = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$

(Discrete) Lyapunov functions

Consider

$$V_k(\xi) = \rho^{-2k} \left(a_0 (f(x_k) - f(x^\star)) + (\xi_k - \xi^\star)^T P (\xi_k - \xi^\star) \right),$$

and assume that we can find $a_0 > 0, P \succeq 0$ such that

$$V_{k+1}(\xi_{k+1}) \leq V_k(\xi_k),$$

we can then conclude

$$f(x_k) - f(x^\star) \leq \rho^{2k} \frac{V_0(\xi_0)}{a_0}.$$

If $\rho < 1$, we have found a convergence rate for $f(x_k)$ towards the optimal value $f(x^\star)$.

A theorem for the (discrete) Lyapunov function

(Discrete) convergence to miminizer

Suppose that there exist $a_0 > 0$, $P \succeq 0$, $\ell > 0$, and $\rho \in [0, 1)$ such that

$$T = M^{(0)} + a_0\rho^2 M^{(1)} + a_0(1-\rho^2)M^{(2)} + \ell M^{(3)} \preceq 0,$$

where

$$M^{(0)} = \begin{bmatrix} A^{\mathcal{T}}PA - \rho^2 P & A^{\mathcal{T}}PB \\ B^{\mathcal{T}}PA & B^{\mathcal{T}}PB \end{bmatrix}, \quad M^{(1)} = N^{(1)} + N^{(2)}, \quad M^{(2)} = N^{(1)} + N^{(3)}, \quad M^{(3)} = N^{(4)},$$

with

$$N^{(1)} = \begin{bmatrix} EA - C & EB \\ 0 & I_d \end{bmatrix}^{\mathcal{T}} \begin{bmatrix} \frac{L}{2}I_d & \frac{1}{2}I_d \\ \frac{1}{2}I_d & 0 \end{bmatrix} \begin{bmatrix} EA - C & EB \\ 0 & I_d \end{bmatrix},$$

$$N^{(2)} = \begin{bmatrix} C - E & 0 \\ 0 & I_d \end{bmatrix}^{\mathcal{T}} \begin{bmatrix} -\frac{m}{2}I_d & \frac{1}{2}I_d \\ \frac{1}{2}I_d & 0 \end{bmatrix} \begin{bmatrix} C - E & 0 \\ 0 & I_d \end{bmatrix},$$

$$N^{(3)} = \begin{bmatrix} C^{\mathcal{T}} & 0 \\ 0 & I_d \end{bmatrix} \begin{bmatrix} -\frac{m}{2}I_d & \frac{1}{2}I_d \\ \frac{1}{2}I_d & 0 \end{bmatrix} \begin{bmatrix} C & 0 \\ 0 & I_d \end{bmatrix},$$

$$N^{(4)} = \begin{bmatrix} C^{\mathcal{T}} & 0 \\ 0 & I_d \end{bmatrix} \begin{bmatrix} -\frac{mL}{m+L}I_d & \frac{1}{2}I_d \\ \frac{1}{2}I_d & -\frac{1}{m+L}I_d \end{bmatrix} \begin{bmatrix} C & 0 \\ 0 & I_d \end{bmatrix}.$$

Then, for $f \in \mathcal{F}_{m,L}$, the sequence $\{x_k\}$ satisfies $f(x_k) - f(x^\star) \le \frac{a_0(f(x_0)-f(x^\star)) + (\xi_0-\xi^\star)^{\mathcal{T}}P(\xi_0-\xi^\star)}{a_0}\rho^{2k}$.

Nesterov method

We introduce $\delta = \sqrt{m\alpha}$ and $d_k = \frac{1}{\delta}(x_k - x_{k-1})$, so we can re-write our algorithm as:

$$\begin{aligned} d_{k+1} &= \beta d_k - \frac{\alpha}{\delta}\nabla f(y_k), \\ x_{k+1} &= x_k + \delta\beta d_k - \alpha\nabla f(y_k), \\ y_k &= x_k + \delta\beta d_k. \end{aligned}$$

Setting $\xi_k = [d_k^{\mathcal{T}}, x_k^{\mathcal{T}}]^{\mathcal{T}} \in \mathbb{R}^{2d}$ we can express the algorithm in the discrete form with

$$A = \begin{bmatrix} \beta I_d & 0 \\ \delta\beta I_d & I_d \end{bmatrix}, \quad B = \begin{bmatrix} -(\alpha/\delta) I_d \\ -\alpha I_d \end{bmatrix}, \quad C = \begin{bmatrix} \delta\beta I_d & I_d \end{bmatrix}, \quad E = \begin{bmatrix} 0 & I_d \end{bmatrix}.$$

THE UNIVERSITY of EDINBURGH

Dimension reduction

- The matrix A is a a Kronecker product of a 2×2 matrix and I_d,

$$A = \begin{bmatrix} \beta & 0 \\ \delta\beta & 1 \end{bmatrix} \otimes I_d;$$

- The matrices B, C and E have a similar Kronecker product structure.
- It is then natural to consider symmetric matrices P of the form

$$P = \widehat{P} \otimes I_d, \qquad \widehat{P} = \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix},$$

- T will also have a Kronecker product structure

$$T = \widehat{T} \otimes I_d, \qquad \widehat{T} = \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ t_{12} & t_{22} & t_{23} \\ t_{13} & t_{23} & t_{33} \end{bmatrix}.$$

Structure of $\widehat{T}$

We have

$$
\begin{aligned}
t_{11} &= \beta^2 p_{11} + 2\delta\beta^2 p_{12} + \delta^2\beta^2 p_{22} - \rho^2 p_{11} - \delta^2\beta^2 m/2, \\
t_{12} &= \beta p_{12} + \delta\beta p_{22} - \rho^2 p_{12} - \delta\beta m/2 + \rho^2\delta\beta m/2, \\
t_{13} &= -\delta^{-1}\alpha\beta p_{11} - 2\alpha\beta p_{12} - \delta\alpha\beta p_{22} + \delta\beta/2, \\
t_{22} &= p_{22} - \rho^2 p_{22} - m/2 + \rho^2 m/2, \\
t_{23} &= -\delta^{-1}\alpha p_{12} - \alpha p_{22} + 1/2 - \rho^2/2, \\
t_{33} &= \delta^{-2}\alpha^2 p_{11} + 2\delta^{-1}\alpha^2 p_{12} + \alpha^2 p_{22} + \alpha^2 L/2 - \alpha.
\end{aligned}
$$

Our task is to find $\rho \in [0,1)$, p_{11}, p_{12}, and p_{22} that lead to $\widehat{T} \preceq 0$ and $\widehat{P} \succeq 0$ (which imply $T \preceq 0$ and $P \succeq 0$).

Solution

The algebra becomes simpler if we represent β and ρ^2 as:

$$\beta = 1 - b\delta, \quad \rho^2 = 1 - r\delta.$$

Note that we are interested in $r \in (0, 1/\delta]$ so as to get $\rho^2 \in [0, 1)$. Going through the algebra we find

$$\widehat{P} = \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} = \frac{m}{2} \begin{bmatrix} (1 - r\delta)^2 & r(1 - r\delta) \\ r(1 - r\delta) & r^2 \end{bmatrix}, \quad \alpha \leq \frac{1}{L}, \quad r \leq 1$$

as well as $\Xi = 0$ where

$$\Xi := \Xi_\delta(r, b) = (r + \delta)(1 - \delta^2)b^2 - 2(1 + r^2)(1 - \delta^2)b + (r^3 - 3r^2\delta + 3r - \delta).$$

- Since $\delta = \sqrt{m\alpha}$ and $\alpha \leq L^{-1}$, this implies that

  $$\rho^2 = 1 - \frac{r}{\sqrt{\kappa}}$$

  hence the Nesterov algorithm maintains the acceleration of the original differential equation.

Convergence of the algorithm

Theorem

With the choices of parameters as in the previous slide the matrix T is negative semi-definite. As a result, for any x_{-1}, x_0, the sequence

$$\rho^{-2k}\left(f(x_k) - f(x_\star) + [d_k^T, x_k^T - x_\star^T]\, P\, [d_k^T, x_k^T - x_\star^T]^T\right)$$

decreases monotonically, which, in particular, implies

$$f(x_k) - f(x_\star) \leq C\rho^{2k}$$

with

$$C = f(x_0) - f(x^\star) + \frac{m}{2}\left\|\frac{1 - r\delta}{\delta}(x_0 - x_{-1}) + r(x_0 - x^\star)\right\|^2.$$

Overview

Connection with the ODE

Convergence between discrete and continuous Lyapunov function

Fix the parameter $\bar{b} > 0$ and the initial conditions $x(0)$, $\dot{x}(0)$ for the momentum equations. For small $h > 0$, consider the Nesterov method with parameters $\alpha = h^2$ and $\beta = \beta_h = 1 - \bar{b}\sqrt{m}h + o(h)$. Assume that the initial points x_{-1}, x_0 are such that, as $h \downarrow 0$, $x_0 \to x(0)$ and $(1/h)(x_0 - x_{-1}) \to \dot{x}(0)$. Then, in the limit $kh \to t$,

1. $x_k \to x(t)$ and $(1/h)(x_{k+1} - x_k) \to \dot{x}(t)$.
2. The discrete Lyapunov function converges to the continuous Lyapunov function

Optimization algorithms as integrators

$$\frac{d}{dt}z = g^{[1]}(z)+g^{[2]}(z)+g^{[3]}(z) := \begin{bmatrix} -\bar{b}\sqrt{m}v \\ 0 \end{bmatrix} + \begin{bmatrix} -\frac{1}{\sqrt{m}}\nabla f(x) \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \sqrt{m}v \end{bmatrix};$$

Nesterov method can be expressed as

$$\begin{aligned}
Z_{k,1} &= z_k, \\
Z_{k,2} &= z_k + hg^{[1]}(Z_{k,1}), \\
Z_{k,3} &= z_k + hg^{[1]}(Z_{k,1}) + hg^{[3]}(Z_{k,2}), \\
Z_{k,4} &= z_k + hg^{[1]}(Z_{k,1}) + hg^{[3]}(Z_{k,2}) + hg^{[2]}(Z_{k,3}), \\
z_{k+1} &= z_k + hg^{[1]}(Z_{k,1}) + hg^{[2]}(Z_{k,3}) + hg^{[3]}(Z_{k,4}).
\end{aligned}$$

Is consistency enough?

1. From an intuitive point of view the previous theorem is obvious, *i.e* you start with and ODE you discretise it and the numerical algorithm inherits its properties for some finite h
2. The key however is how large this h can be, while maintaining the negative definiteness of the matrix T.
3. From consistency in order to achieve acceleration one needs to be able to preserve the negative definiteness of T for time steps $h \leq cL^{-1/2}$
4. What is special about Nesterov?

THE UNIVERSITY of EDINBURGH

Structural conditions of integrators

$$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha \nabla f(y_k),$$
$$y_k = x_k + \gamma(x_k - x_{k-1}),$$

- Key quantity $c := t_{11}/(m\delta)$, when $\gamma = 0$, $c = \cdots + \delta(\kappa - 1)\beta^2/2$.
- For acceleration, δ has to be $\mathcal{O}(1/\sqrt{\kappa})$ which makes it impossible for c to be ≤ 0.
- Presence of κ in t_{11} relates to the appearance of L in the matrix $N^{(1)}$
- This can be indeed eliminated if $EA - C = 0$
- In words: the point $y_k = C\xi_k$ where the gradient is evaluated has to coincide with the point $x_{k+1} = EA\xi_k$ that the algorithm would yield if $u_k = \nabla f(y_k)$ happened to vanish

[3] L. Lessard, B. Recht, A. Packard, *SIAM J. Optim.*, 26(1), 57–95. (2016)

Revisiting the Lyapunov function

$$V(\xi, t) = e^{\lambda t}\left(f(y(t)) - f(y^\star) + (\xi(t) - \xi^\star)^{\mathcal{T}}\bar{P}(\xi(t) - \xi^\star)\right)$$

- We can try to relax the condition $\bar{P} \succeq 0$
- Through strong convexity we know that

$$f(y(t)) - f(y^\star) \geq \frac{m}{2}\|y(t) - y^\star\|^2.$$

- Hence

$$V(\xi, t) \geq e^{\lambda t}\left[(\xi(t) - \xi^\star)^{\mathcal{T}}\left(\frac{m}{2}\bar{C}^T\bar{C} + \bar{P}\right)(\xi(t) - \xi^\star)\right]$$

- If we can still establish that $V(\xi, t)$ is non-increasing we are good as long $\bar{C}^T\bar{C} + \bar{P} \succeq 0$

THE UNIVERSITY of EDINBURGH

Continuous case revisited

Improved (continuous) convergence to minimizer

Suppose that there exist $\lambda > 0$, $\sigma \geq 0$ and a symmetric matrix $\bar{P}$ with $\widetilde{P} := \bar{P} + (m/2)\bar{C}^T\bar{C} \succ 0$, that satisfy

$$\bar{T} = \bar{M}^{(0)} + \bar{M}^{(1)} + \lambda\bar{M}^{(2)} + \sigma\bar{M}^{(3)} \preceq 0$$

Then the following inequality holds for $f \in \mathcal{F}_{m,L}$, $t \geq 0$

$$\|y(t) - y_*\|^2 \leq \max\sigma(\bar{C}^T\bar{C})\,\|\xi(t) - \xi^*\|_{\widetilde{P}} \leq \frac{\max\sigma(C^TC)}{\min\sigma(\widetilde{P})} e^{-\lambda t} V(\xi(0), 0).$$

Discrete case revisited

Improved (discrete) convergence to minimizer

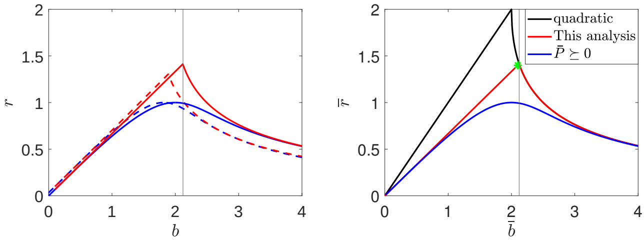
Suppose that there exist $a_0 > 0$, $\rho \in (0,1)$, $\ell > 0$, and a symmetric matrix P, with $\widetilde{P} := P + (a_0 m/2)E^T E \succ 0$, such that

$$T = M^{(0)} + a_0\rho^2 M^{(1)} + a_0(1-\rho^2)M^{(2)} + \ell M^{(3)} \preceq 0,$$

Then, for $f \in \mathcal{F}_{m,L}$, the sequence $\{x_k\}$ satisfies

$$\|x_k - x_\star\|^2 \leq \max\sigma(E^T E)\,\|\xi_k - \xi^\star\|_{\widetilde{P}} \leq \frac{\max\sigma(E^T E)}{\min\sigma(\widetilde{P})} V(\xi_0, 0)\rho^{2k}.$$

What do we gain?



- We can show that in continuous time for $\bar{b} = 3\sqrt{2}/2$ we can improve the convergence rate to $\lambda = \sqrt{2}\sqrt{m}$
- In the discrete setting for appropriate choice of the coefficients we can prove a convergence rate $\rho^2 = 1 - \frac{\sqrt{2}}{\sqrt{\kappa}} + \mathcal{O}(\kappa^{-1}), \quad \kappa \to \infty$.
- The convergence rate of Nesterov with the standard parameter choices $\alpha = L^{-1}, \beta = (\sqrt{k} - 1)/(\sqrt{k} + 1)$ is better that what previously proven.

Overview

THE UNIVERSITY of EDINBURGH

Conclusions

- Differential equations are excellent starting point in terms of designing optimization algorithms.
- However for optimization algorithms stability is crucial in terms of being able to utilize the favourable convergence rates of the continuous system.
- In terms of Lyapunov functions it is possible to improve on previous convergence rates by relaxing some conditions by using the strong convexity properties of our functions.

Bibliography

1. M. Fazlyab, A. Ribeiro, M. Morari, and V. M. Preciado. Analysis of optimization algorithms via integral quadratic constraints: nonstrongly convex problems. *SIAM Journal on Optimization*, 28(3):2654–2689, 2018
2. W. Su, S. Boyd, and E. J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.
3. L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.