Matrix stabilization using differential equations.

Nicola Guglielmi

Universitá dell'Aquila and Gran Sasso Science Institute, Italia

NUMOC-2017 Roma, 19–23 June, 2017

Inspired by a joint work with Christian Lubich (Tübingen).

Problem statement

For a given unstable matrix $A \in \mathbb{S}$ with $\mathbb{S}$ a prescribed structure, stabilization consists in looking for a nearest stable matrix $B \in \mathbb{S}$.

Few extensions

- Feedback stabilization of a linear control system (a classical open problem in control theory);
- Stabilization of polynomials (through companion matrices);
- Stabilization of gyroscopic systems;
- Computation of the closest correlation matrix (symmetric positive definite with unit diagonal) to a symmetric matrix (relevant problem from finance, see e.g. **N. Higham** webpage).

This distance is measured in a matrix norm, often the spectral or the Frobenius norm. In this talk we consider the Frobenius norm.

Feedback stabilization of linear control systems

Consider the linear dynamical system with input and output defined by

$$\begin{cases} \dot{x}(t) & = & A\,x(t) + B\,u(t), \\ y(t) & = & C\,x(t) \end{cases}$$

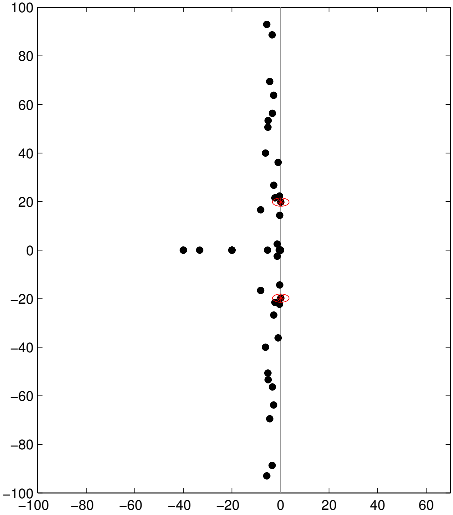
where $A \in \mathbb{R}^{n,n}$, $B \in \mathbb{R}^{n,p}$, $C \in \mathbb{R}^{q,n}$, x is the state, u is the input and y is the output. Setting the control u proportional to the output, $u = \Delta y$ (with $\Delta \in \mathbb{C}^{p,q}$) one gets the ODE

$$\dot{x}(t) = (A + B\Delta C)\,x(t).$$

where $\Delta \in \mathbb{C}^{p,q}$. If A is not Hurwitz, a fundamental problem is to find a stabilizing feedback, i.e. a matrix Δ of minimum norm s.t. $A + B\Delta C$ is Hurwitz.

The constraints on actuators require Δ as small as possible.

Example: Boeing767-matrix



From flutter analysis of the Boeing 767 aircraft. A unstable (two eigenvalues in $\mathbb{C}^+$) B and C are fixed $n \times 2$ and $2 \times n$ matrices.

Stabilization: motivation and preliminary ideas

For a linear system of ODEs, arising from discretization of a PDE and by successive model reduction, it is possible that peculiar properties, like stability are lost. For this reason it is important to provide a minimal norm correction of the matrix A, say $A + \Delta$, which is stable.

How to stabilize

A natural idea is that of sweeping the eigenvalues in the right complex plane to the left, along an optimal path traveled by the eigenvalues in the complex plane under a smoothly varying perturbation to A.

Drawback of individual sweeping

This however turns out to be very difficult and often inefficient. The reason is that at the same time several eigenvalues from the left usually move to the right and controlling a large part of the spectrum woud be very demanding.

Outline of the talk

1. **Framework**
2. **Two step methodology**
3. **Inner step**
   - Derivation of monotone ODEs
   - Low-rank ODEs
4. **Outer step**
   - Fast approximation of the distance
5. **Large size examples**

Stabilization of an unstable system

Given $A \in \mathbb{S}$ consider $y' = Ay$ with eigenvalues with positive real part.

- How far is A from a Hurwitz matrix?
- For ε s.t. $\|\Delta\| \le \varepsilon$, how small can be spectral abscissa of $A + \Delta$?

This is a difficult optimization problem, it is non-convex and non-smooth (the spectral abscissa is non-Lipschitz).

Some recent literature

- **Nesterov, Orbandexivry and van Dooren, 2013** propose an iterative method based on successive convex approximations. The method behaves well but its efficiency is limited by the dimension.
- **Overton, 2012** proposes a BFGS-based penalization method

  $$f(\Delta) = \|\Delta|\|_F + \rho\, \alpha(A + \Delta) \qquad \text{(with } \rho > 0 \text{ if } \alpha(A+\Delta) > 0\text{)}$$

  In our experiments results sometimes far from optimality.
- **Gillis & Sharma, 2017** propose a reformulation using linear dissipative Hamiltonian systems. Very good results for small problems.

Problem setting

Let $A \in \mathbb{S}$, where $\mathbb{S}$ denotes a structured set of matrices e.g.

$\mathbb{S} = \mathbb{R}^{n,n}$, matrices with prescribed sparsity pattern, Toeplitz ...

an unstable matrix, having few eigenvalues in the right complex plane.

A matrix nearness problem. Compute

$$d^{\mathbb{S}}(A) = \inf\Big\{ \|A - B\| : \, B \in \mathbb{S} \text{ is Hurwitz} \Big\}.$$

One clear fact: the matrix B has necessarily eigenvalues on imaginary axis. If not, by continuity of eigenvalues $\|A - B\|$ may be reduced.

For practical purposes it seems natural to prefer a structured distance although the unstructured one has also interest.

Synthesis of our contribute

Methodology

We use a two-level procedure that uses matrix ODEs on the inner level and a fast one-dimensional optimization on the outer level.

Local optimality

As with the algorithms proposed in the literature, our method is not guaranteed to yield the perturbation of globally minimal norm. However, our method is globally convergent to a local optimum.

Numerical experience

In our many numerical experiments we found, however, that the approach presented here yields stabilizing perturbations that have a norm comparable to or smaller than those given by the algorithm of **Nesterov, Orbandexivry & van Dooren**, at significantly reduced computational cost especially for matrices of dimension $d \geq 30$.

Outline of the talk

1 Framework

2 **Two step methodology**

3 Inner step
- Derivation of monotone ODEs
- Low-rank ODEs

4 Outer step
- Fast approximation of the distance

5 Large size examples

Sketch of the methodology

Notation. Let m_+ denote the number of eigenvalues with positive real part of the given matrix A.

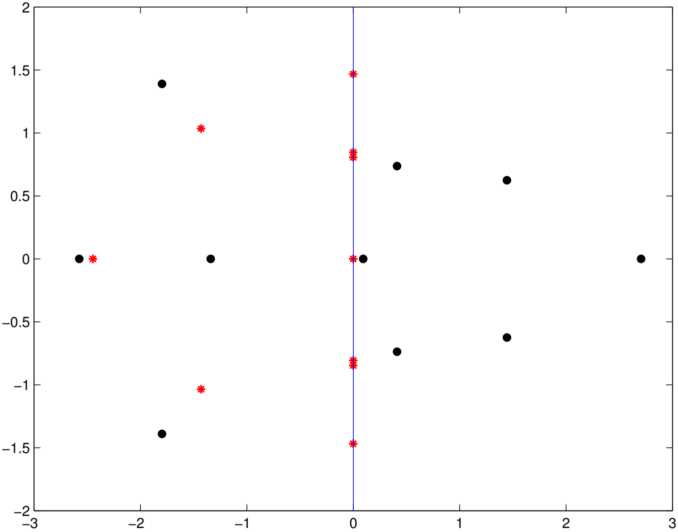
Our approach can be sketched as follows.

- Choose an integer m (usually $m \geq m_+$) and consider the m eigenvalues with largest real part of perturbations $A + \Delta$.
- Find a perturbation Δ of minimal norm such that these m eigenvalues are aligned on the imaginary axis.

Remark: for $m > m_+$, it may appear counterintuitive to try to align more eigenvalues on the imaginary axis than the original number m_+.

- If A is Hermitian this result is rigorous (by Weyl theory).
- In the general case this is motivated by numerical results with the algorithm of **Nesterov, Orbandexivry & van Dooren** where such an alignment is observed in the stabilized matrix.

Example: a 10×10 matrix, $m = 6$



in black the eigenvalues of A; in red those of $A + \Delta$.

A measure of instability

For a given m and a fixed perturbation size $\varepsilon > 0$, we write $\Delta = \varepsilon E$ (with $\|E\|_F = 1$) and minimize the function

$$F_\varepsilon(E) = \frac{1}{2}\sum_{i=1}^{m}\Big(\mathrm{Re}\ \lambda_i(A+\varepsilon E)\Big)^2 \qquad \text{(constrained by } \|E\|_F = 1\text{)},$$

where $\{\lambda_i\}_{i=1}^m$ have largest real part. If $F_\varepsilon(E) > 0$ we have a measure of minimal instability associated to an optimal perturbation of norm ε.

Remark: we prove that the optimal perturbation matrix E has rank at most m. This motivates to formulate an algorithm that considers only rank-m perturbations. Interesting when $n \gg m$.

Modified functional

Instead of the functional F_ε, one could choose a different functional

$$\Phi_\varepsilon(E) = \sum_{i=1}^{m} \omega_i \left(\text{Re } \lambda_i(A + \varepsilon E)\right)$$

with nonnegative weights ω_i, which may depend on E.

A natural choice is to choose ω_i proportional to Re $\lambda_i(A + \varepsilon E)$ for those i where the real part is larger than 0. So one might choose

$$\omega_i = \max\left\{0, \text{Re } \lambda_i(A + \varepsilon E)\right\}.$$

For the following we have chosen to work with the smooth functional F_ε, but we note that the whole program could be carried out also for Φ_ε with minor modifications.

Two-level procedure for matrix matrix stabilization

Assume $\mathbb{S}$ is a linear space of matrices.

Matrix A is given.

(i) For given $\varepsilon > 0$, find $E = E(\varepsilon) \in \mathbb{S}$ of unit norm which minimizes $F_\varepsilon(E)$.

(ii) Modify ε until $F_\varepsilon(E) = 0$ i.e. find $\varepsilon^* \longrightarrow \min_{\varepsilon>0}\Big\{F_\varepsilon(E(\varepsilon)) = 0\Big\}$.

---

use a structured ODE for determining E in the first step

Outline of the talk

1. Framework
2. Two step methodology
3. Inner step
   - Derivation of monotone ODEs
   - Low-rank ODEs
4. Outer step
   - Fast approximation of the distance
5. Large size examples

An ODE approach

Idea: to construct a smooth matrix valued function (here ε is fixed)

$$A + \varepsilon E(t), \qquad \text{with } E(t) \in \mathbb{S} \text{ and } \|E(t)\|_F \equiv 1$$

such that $F_\varepsilon(E(t))$ is decreasing w.r.t. t.

Lemma (Basic perturbation result: Wilkinson, Kato, ...)

Let $t \mapsto E(t)$ be a differentiable matrix valued function, and $\lambda(t)$ a path of simple eigenvalues of $A + \varepsilon E(t)$. Then,

$$\dot\lambda(t) = \frac{y(t)^* \dot E(t) x(t)}{y(t)^* x(t)} = \sigma(t) \left\langle y(t)x(t)^*, \dot E(t) \right\rangle$$

where $y(t)$, $x(t)$ are left and right eigenvectors of $A + \varepsilon E(t)$ to $\lambda(t)$ (in the sequel we assume $\sigma(t) = 1/y(t)^* x(t) > 0$).

$\langle A, B \rangle = \mathrm{trace}\,(A^* B)$ denotes the Frobenius inner product

Free gradient

Consider $F_\varepsilon(E) : \mathbb{C}^{n,n} \longrightarrow \mathbb{R}$

Let $\dot{E}(t)$ the derivative of a smooth matrix-valued function $E(t)$, then

$$\frac{d}{dt} F_\varepsilon \left(E(t)\right) = \varepsilon \, \mathrm{Re} \left\langle \sum_{i=1}^{m} \gamma_i(t) \, y_i(t) x_i(t)^*, \dot{E}(t) \right\rangle$$

with $\gamma_i(t) = \mathrm{Re}\ \lambda_i\big(A + \varepsilon E(t)\big) / y_i(t)^* x_i(t), \quad i = 1, \ldots, m.$

This yields the free gradient

$$G_0(E) = \sum_{i=1}^{m} \gamma_i y_i x_i^* \qquad \text{a rank-}m \text{ matrix}$$

Projected gradient of the functional

The optimization problem

$$Z_* = \arg\min_{Z \in \mathbb{S}} \ \mathrm{Re} \left\langle \sum_{i=1}^{m} \gamma_i \, y_i x_i^*, Z \right\rangle$$

$$\text{subj to} \quad \mathrm{Re}\,\langle E, Z \rangle = 0 \quad \text{(norm preservation)}$$

$$\text{and} \quad \|Z\|_F = 1 \quad \text{(for uniqueness)}$$

The solution is given by

$$Z_* \propto \, - G(E) = \, - P_{\mathbb{S}}\big(G_0(E)\big) + \mu E,$$

where $P_{\mathbb{S}}(\cdot)$ denotes orthogonal projection onto $\mathbb{S}$ and $G(E)$ the projected gradient onto manifold $\mathbb{S} \cap \{Z \in \mathbb{C}^{n,n} : \mathrm{Re}\,\langle E, Z \rangle = 0\}$.

Gradient system

In order to minimize F_ε we consider the ODE $\dot{E} = -G(E)$,

$$\dot{E} = -P_{\mathbb{S}}\left(G_0(E)\right) + \mu E, \qquad \text{with} \quad \mu = \mathrm{Re}\left\langle P_{\mathbb{S}}\left(G_0(E)\right), E\right\rangle$$

Theorem

The flow of the ODE has the following properties:

1. *Norm conservation:* $\|E(t)\|_F = 1$ *for all* t*;*
2. *Monotonicity:* $F_\varepsilon\left(E(t)\right)$ *decreasing along solutions of ODE;*
3. *Stationary points: the following statements are equivalent:*

$$\frac{d}{dt}F_\varepsilon\big(E(t)\big) = 0 \iff \dot{E} = 0 \iff E \text{ real multiple of } P_{\mathbb{S}}\left(G_0(E)\right)$$

Issues: Computing stationary points of the ODEs efficiently.

Low-rank ODEs in the case $\mathbb{S} = \mathbb{C}^{n,n}$

Note: Stationary points are obtained by

$$\dot{E} = -G_0(E) + \mu E = 0 \quad \Longrightarrow \quad E \propto G_0(E) = \sum_{i=1}^{m} \gamma_i y_i x_i^* \quad \text{(rank-}m\text{)}.$$

Alas the solution of the ODE has full rank (even if $E(0)$ has rank-m).

Low-rank ODEs in the case $\mathbb{S} = \mathbb{C}^{n,n}$

Note: Stationary points are obtained by

$$\dot{E} = -G_0(E) + \mu E = 0 \quad \Longrightarrow \quad E \propto G_0(E) = \sum_{i=1}^{m} \gamma_i y_i x_i^* \quad \text{(rank-}m\text{)}.$$

Alas the solution of the ODE has full rank (even if $E(0)$ has rank-m).

Main result: dynamics on the manifold $\mathcal{M}^m$ of rank-m matrices by F-orthogonal projection Π_E^m to tangent space $T_E\mathcal{M}^m$:

$$\dot{E} = -\Pi_E^m\big(-G_0(E) + \mu E\big)$$

Theorem

The projected ODE has the same properties of the unprojected ODE (monotonicity, norm conservation, stationary points).

Rank-m differential equations

Every rank-m matrix E can be written in the form $E = USV^*$ where $U, V \in \mathbb{C}^{n,m}$ are orthonormal and $S \in \mathbb{C}^{m,m}$ is invertible.

The decomposition is not unique but we use a unique decomposition in the tangent space. Using the explicit projection formula

$$\Pi_E^m(G_0) = G_0VV^* - UU^*G_0VV^* + UU^*G_0, \qquad G_0 \in \mathbb{C}^{n,n}$$

we get for the projected ODE

$$\begin{cases} \dot{U} &= (I - UU^*)G_0VS^{-1} \quad (U \in \mathbb{C}^{n,m} \text{ orthonormal}) \\ \dot{V} &= (I - VV^*)G_0^*US^{-*} \quad (V \in \mathbb{C}^{n,m} \text{ orthonormal}) \\ \dot{S} &= U^*G_0V \qquad\qquad\quad (S \in \mathbb{C}^{m,m} \text{ of unit norm}) \end{cases}$$

Implementation

Difficulties: the right-hand sides of the ODEs for U and V have S^{-1}, which may cause numerical difficulties with standard integrators when S is nearly singular, i.e. E is close to a matrix of rank smaller than m.

1. We follow the approach by **Lubich & Oseledets, 2014**, based on splitting the tangent space projection Π_E^m, that is an alternating sum of three subprojections. A time step is based on the Lie–Trotter splitting corresponding to these three terms.
2. Step size control based on decrease of $F_\varepsilon(E)$ (no need to follow the trajectory accurately)
3. Eigenvalue computation by implicitly restarted Arnoldi method exploiting possibly sparse plus low-rank structure of $A + \varepsilon E$.

Outline of the talk

1 **Framework**

2 **Two step methodology**

3 **Inner step**
- Derivation of monotone ODEs
- Low-rank ODEs

4 **Outer step**
- Fast approximation of the distance

5 **Large size examples**

A key variational formula

Given ε we compute a local extremizer $E(\varepsilon)$ which minimizes $F_\varepsilon(E)$.
In order to approximate the distance we have to solve equation

$F_\varepsilon(E(\varepsilon)) = 0$ with respect to ε.

Derivative of F_ε

Theorem

Let $E(\varepsilon)$ be a smooth path of extremizers s.t. $F_\varepsilon(E(\varepsilon)) > 0$. Then

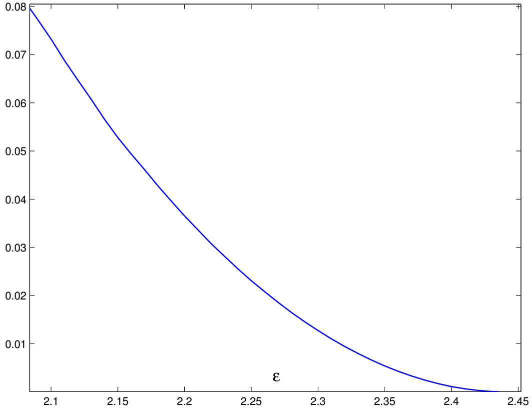
$$\frac{dF_\varepsilon(E(\varepsilon))}{d\varepsilon} = -\|G_0(\varepsilon)\|_F < 0$$

with $G_0(\varepsilon) = \sum_{i=1}^{m} \gamma_i(\varepsilon)\, y_i(\varepsilon) x_i(\varepsilon)^*$.

Optimal stabilization

We denote as ε^* the smallest root of $F_\varepsilon(E(\varepsilon)) = 0$.

The zero ε^* of $F_\varepsilon(E(\varepsilon))$ is generically double and we use a modified Newton iteration which converges quadratically from the left. From the right we use bisection which provides linear reduction of the error.



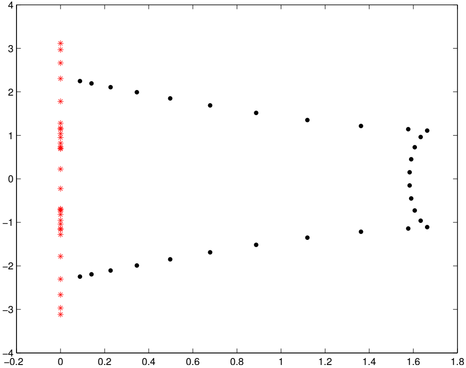
An interesting small example: the 30×30 Grcar matrix.

Nesterov et al. gives
$\|B - A\|_F \approx 6.50$ in > 48 hours

BFGS method gives
$\|B - A\|_F \approx 27.01$ in 79 seconds

The ODE method gives
$\|B - A\|_F \approx 6.57$ in 143 seconds

Gillis-Sharma method improves to
$\|B - A\|_F \approx 6.11$ in 101 seconds
Quite disappointing!



Spectra of A and B (ODE method)

Peculiarity of the ODE approach The spectral abscissa is not smooth due to rightmost eigenvalues exchange. An ODE approach sweeping all the eigenvalues in $\mathbb{C}^+$ can smooth out the problem.

Outline of the talk

1 Framework

2 Two step methodology

3 Inner step
- Derivation of monotone ODEs
- Low-rank ODEs

4 Outer step
- Fast approximation of the distance

5 **Large size examples**

Large size example: Brusselator

This sparse matrix[1] (of size = 800) arises from form a 2-dimensional reaction-diffusion model in chemical engeenering.

Algorithm	Structure	Norm of Stabilizer Δ
GL	$\mathbb{R}$	0.07750
GS	$\mathbb{R}$	0.15102
Ov	$\mathbb{R}$	0.15102
ONV	$\mathbb{R}$	not computed

Legenda. The acronym GL stands for Guglielmi-Lubich, GS for Gillis-Sharma, Ov for Overton and ONV for Orbandexivry, Nesterov, and Van Dooren.

[1] see http://math.nist.gov/MatrixMarket/data/NEP/brussel/brussel.html

Large size example: Brusselator...

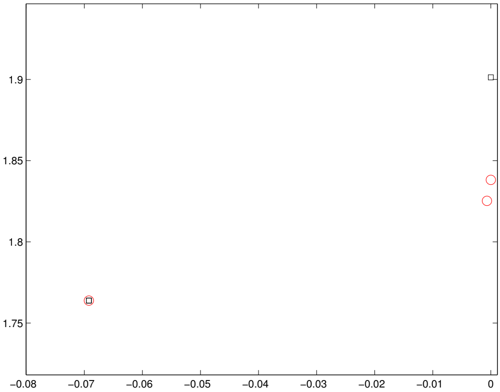


Figure: Rightmost eigenvalues of the stabilized matrix computed by the methods Ov and GS (black squares) and by the GL method (red circles). The eigenvalue with real part close to -0.07 is double for the stabilized matrices computed by the GS and Ov methods while it is simple for the stabilized matrix computed by the GL method. Indeed moving the double eigenvalue to the right would allow to reduce the norm of the stabilized matrix in the methods by GS and Ov.

Large size example: Tolosa

Tolosa[2] matrix (size = 1090, type sparse) arises in stability analysis of a model of an airplane in flight. Indeeed the shifted matrix $A + \frac{1}{2}\mathrm{I}$ is considered here, in order to robustly increase stability of the matrix.

Algorithm	**Structure**	**Norm of Stabilizer Δ**
GL	$\mathbb{R}$	157.930
GS	$\mathbb{R}$	287.957 (∞)
Ov	$\mathbb{R}$	$6.0131 \cdot 10^6$
ONV	$\mathbb{R}$	not computed

[2]see http://math.nist.gov/MatrixMarket/data/NEP/mvmtls/mvmtls.html

Some issues

- Robustness of the method is increased by varying m;
- Characterization of globally optimal solutions of the stabilization problem is hard and open;
- Use of 2-norm and other norms unexplored;
- Some extensions of the proposed approach are very natural but require technical developments;
- Adaptation to related problems and exploitation of the underlying low-rank structure is also to be studied (for example when imposing preservation of sparsity pattern);
- Improving convergence speed to stationary points of the ODEs.