On projection methods for large-scale Riccati equations

V. Simoncini

Dipartimento di Matematica

Alma Mater Studiorum - Università di Bologna

valeria.simoncini@unibo.it

The problem

Find $\mathbf{X} \in \mathbb{R}^{n\times n}$ such that

$$A\mathbf{X} + \mathbf{X}A^\top - \mathbf{X}BB^\top\mathbf{X} + C^\top C = 0$$

with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times p}$, $C \in \mathbb{R}^{s\times n}$, $p, s = \mathcal{O}(1)$

Rich literature on analysis, applications and numerics:

Lancaster-Rodman 1995, Bini-Iannazzo-Meini 2012, Mehrmann etal 2003 ...

The problem

Find $\mathbf{X} \in \mathbb{R}^{n\times n}$ such that

$$A\mathbf{X} + \mathbf{X}A^\top - \mathbf{X}BB^\top\mathbf{X} + C^\top C = 0$$

with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times p}$, $C \in \mathbb{R}^{s\times n}$, $p, s = \mathcal{O}(1)$

Rich literature on analysis, applications and numerics:

Lancaster-Rodman 1995, Bini-Iannazzo-Meini 2012, Mehrmann etal 2003 ...

We focus on the large scale case: $n \gg 1000$

Different strategies

- (Inexact) Kleinman iteration (Newton-type method)
- Projection methods
- Invariant subspace iteration
- (Sparse) multilevel methods
- ....

Recent Comparisons: Simoncini-Szyld-Monsalve, '14, Benner etal, '18, etc.

Galerkin projection method for the Riccati equation

Given the orth basis V_k for an approximation space, determine

$$\mathbf{X}_k = V_k \mathbf{Y}_k V_k^\top$$

to the Riccati solution matrix by orthogonal projection:

Galerkin condition:	Residual orthogonal to approximation space

Galerkin projection method for the Riccati equation

Given the orth basis V_k for an approximation space, determine

$$\mathbf{X}_k = V_k \mathbf{Y}_k V_k^\top$$

to the Riccati solution matrix by orthogonal projection:

Galerkin condition:	Residual orthogonal to approximation space

$$V_k^\top (A\mathbf{X}_k + \mathbf{X}_k A^\top - \mathbf{X}_k B B^\top \mathbf{X}_k + C^\top C) V_k = 0$$

Galerkin projection method for the Riccati equation

Given the orth basis V_k for an approximation space, determine

$$\mathbf{X}_k = V_k \mathbf{Y}_k V_k^\top$$

to the Riccati solution matrix by orthogonal projection:

Galerkin condition:	Residual orthogonal to approximation space

$$V_k^\top (A\mathbf{X}_k + \mathbf{X}_k A^\top - \mathbf{X}_k BB^\top \mathbf{X}_k + C^\top C)V_k = 0$$

giving the reduced Riccati equation

$$(V_k^\top A V_k)\mathbf{Y} + \mathbf{Y}(V_k^\top A^\top V_k) - \mathbf{Y}(V_k^\top BB^\top V_k)\mathbf{Y} + (V_k^\top C^\top)(CV_k) = 0$$

$\mathbf{Y}_k$ is the stabilizing solution (Jaimoukha-Kasenally 1994, Heyouni-Jbilou 2009)

Galerkin projection method for the Riccati equation

Given the orth basis V_k for an approximation space, determine

$$\mathbf{X}_k = V_k \mathbf{Y}_k V_k^\top$$

to the Riccati solution matrix by orthogonal projection:

Galerkin condition:	Residual orthogonal to approximation space

$$V_k^\top (A\mathbf{X}_k + \mathbf{X}_k A^\top - \mathbf{X}_k B B^\top \mathbf{X}_k + C^\top C) V_k = 0$$

giving the reduced Riccati equation

$$(V_k^\top A V_k)\mathbf{Y} + \mathbf{Y}(V_k^\top A^\top V_k) - \mathbf{Y}(V_k^\top B B^\top V_k)\mathbf{Y} + (V_k^\top C^\top)(C V_k) = 0$$

$\mathbf{Y}_k$ is the stabilizing solution (Jaimoukha-Kasenally 1994, Heyouni-Jbilou 2009)

Key questions:

- Which approximation space?
- Is this meaningful from a Control Theory perspective?

Dynamical systems and the Riccati equation

Time-invariant linear system

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \qquad x(0) = x_0 \\ y(t) = Cx(t), \end{cases}$$

$u(t)$: control (input) vector; $y(t)$: output vector

$x(t)$: state vector; x_0 : initial state

Dynamical systems and the Riccati equation

Time-invariant linear system

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \qquad x(0) = x_0 \\ y(t) = Cx(t), \end{cases}$$

$u(t)$: control (input) vector; $y(t)$: output vector

$x(t)$: state vector; x_0 : initial state

Minimization problem for a Cost functional: (simplified form)

$$\inf_u \mathcal{J}(u, x_0) \qquad \mathcal{J}(u, x_0) := \int_0^\infty \big(x(t)^\top C^\top C x(t) + u(t)^\top u(t)\big)\, dt$$

Dynamical systems and the Riccati equation

$$\inf_u \mathcal{J}(u, x_0) \qquad \mathcal{J}(u, x_0) := \int_0^\infty \left(x(t)^\top C^\top C x(t) + u(t)^\top u(t) \right) dt$$

Riccati equation:

$$A^\top \mathbf{X} + \mathbf{X} A - \mathbf{X} B B^\top \mathbf{X} + C^\top C = 0$$

THEOREM. Let the pair (A, B) be stabilizable and (C, A) observable. Then there is a unique solution $\mathbf{X} \geq 0$ of the Riccati equation. Moreover,

i) For each x_0 there is a unique optimal control, and it is given by

$$u_*(t) = -B^\top \mathbf{X} \exp((A - BB^\top \mathbf{X})t) x_0 \quad \text{for} \quad t \geq 0$$

ii) $\mathcal{J}(u_*, x_0) = x_0^\top \mathbf{X} x_0$ for all $x_0 \in \mathbb{R}^n$

see, e.g., Lancaster & Rodman, 1995

Order reduction of dynamical systems by Galerkin projection

Let $V_k \in \mathbb{R}^{n \times d_k}$ have orthonormal columns, $d_k \ll n$

Let $\quad T_k = V_k^\top A V_k, \quad B_k = V_k^\top B, \quad C_k^\top = V_k^\top C^\top$

Reduced order dynamical system:

$$\begin{cases} \dot{\widehat{x}}(t) = T_k \widehat{x}(t) + B_k \widehat{u}(t), & \widehat{x}(0) = \widehat{x}_0 := V_k^\top x_0 \\ \widehat{y}(t) = C_k \widehat{x}(t) \end{cases}$$

$$\boxed{x_k(t) = V_k \widehat{x}(t) \approx x(t)}$$

Typical frameworks:

- Transfer function approximation
- Model reduction

⋆ Petrov-Galerkin projection is also common (see, e.g., Antoulas '05)

Reduced Riccati equation

$$T_k^\top \mathbf{Y} + \mathbf{Y} T_k - \mathbf{Y} B_k B_k^\top \mathbf{Y} + C_k^\top C_k = 0 \qquad (*)$$

THEOREM. Let the pair (T_k, B_k) be stabilizable and (C_k, T_k) observable. Then there is a unique solution $\mathbf{Y}_k \geq 0$ of $(*)$ that for each $\widehat{x}_0$ gives the feedback optimal control

$$\widehat{u}_*(t) = -B_k^* \mathbf{Y}_k \exp((T_k - B_k B_k^* \mathbf{Y}_k)t)\widehat{x}_0, \quad t \geq 0$$

for the reduced system.

Reduced Riccati equation

$$T_k^\top \mathbf{Y} + \mathbf{Y} T_k - \mathbf{Y} B_k B_k^\top \mathbf{Y} + C_k^\top C_k = 0 \qquad (*)$$

THEOREM. Let the pair (T_k, B_k) be stabilizable and (C_k, T_k) observable. Then there is a unique solution $\mathbf{Y}_k \geq 0$ of $(*)$ that for each $\widehat{x}_0$ gives the feedback optimal control

$$\widehat{u}_*(t) = -B_k^* \mathbf{Y}_k \exp((T_k - B_k B_k^* \mathbf{Y}_k)t)\widehat{x}_0, \quad t \geq 0$$

for the reduced system.

♣ If there exists a matrix K such that $A - BK$ is dissipative[a], then the pair (T_k, B_k) is stabilizable.

[a] A matrix is dissipative if its field of values is all in $\mathbb{C}^-$.

Reduced optimal control vs approximate control

⋆ Our reduced optimal control function:

$$\widehat{u}_*(t) = -B_k^\top \mathbf{Y}_k e^{(T_k - B_k B_k^\top \mathbf{Y}_k)t} \widehat{x}_0, \quad t \geq 0$$

⋆ Commonly used approximate control function:

Consider the Riccati equation

$$A^\top \mathbf{X} + \mathbf{X}A - \mathbf{X}BB^\top \mathbf{X} + C^\top C = 0$$

If $\widetilde{\mathbf{X}}$ is some approximation to $\mathbf{X}$, then

$$\widetilde{u}(t) := -B^\top \widetilde{\mathbf{X}} \widetilde{x}(t) \quad \text{where} \quad \widetilde{x}(t) := e^{(A - BB^\top \widetilde{\mathbf{X}})t} x_0$$

However,

$$\boxed{\widehat{u}_* \neq \widetilde{u}}$$

They induce different actions on the functional $\mathcal{J}$ (even for $\widetilde{\mathbf{X}} \equiv V_k \mathbf{Y}_k V_k^\top$)

Reduced optimal control vs approximate control

Consider the interpolated approximation: $\mathbf{X}_k = V_k \mathbf{Y}_k V_k^\top$

Riccati **residual** matrix:

$R_k := A^\top \mathbf{X}_k + \mathbf{X}_k A - \mathbf{X}_k B B^\top \mathbf{X}_k + C^\top C$

⋆ Reduced optimal control function: $\widehat{u}_*(t) = -B_k^\top \mathbf{Y}_k e^{(T_k - B_k B_k^\top \mathbf{Y}_k)t} \widehat{x}_0$

THEOREM. Assume that $A - BB^\top \mathbf{X}_k$ is stable and $\widetilde{u}(t) := -B^\top \mathbf{X}_k x(t)$ approx control. Then

$$|\mathcal{J}(\widetilde{u}, x_0) - \widehat{\mathcal{J}}_k(\widehat{u}_*, \widehat{x}_0)| = \mathcal{E}_k, \quad \text{with} \quad \mathcal{E}_k \le \frac{\|R_k\|}{2\alpha} x_0^\top x_0,$$

where $\alpha > 0$ is such that $\|e^{(A - BB^\top \mathbf{X}_k)t}\| \le e^{-\alpha t}$ for all $t \ge 0$.

Note: $|\mathcal{J}(\widetilde{u}, x_0) - \widehat{\mathcal{J}}_k(\widehat{u}_*, \widehat{x}_0)|$ is nonzero for $R_k \neq 0$

On the choice of the reduction space

Reduced problem, $T_k = V_k^\top A V_k$, $B_k = V_k^\top B$, $C V_k = C_k$,

$$T_k^\top \mathbf{Y}_k + \mathbf{Y}_k T_k - \mathbf{Y}_k B_k B_k^\top \mathbf{Y}_k + C_k{}^\top C_k = 0$$

$\mathcal{K} = \mathrm{Range}(V_k)$:

♣ Krylov-type subspaces (extensively used in the linear case)

- $\mathcal{K}_k(A, C^\top) := \mathrm{Range}([C^\top, AC^\top, \ldots, A^{k-1}C^\top])$ (Polynomial)
- $\mathcal{EK}_k(A, C^\top) := \mathcal{K}_k(A, C^\top) + \mathcal{K}_k(A^{-1}, A^{-1}C^\top)$ (EKS, Rational)
- $\mathcal{RK}_k(A, C^\top, \mathbf{s}) := \mathrm{Range}([C^\top, (A - s_2 I)^{-1}C^\top, \ldots, \prod_{j=1}^{k-1}(A - s_{j+1}I)^{-1}C^\top])$

  (RKS, Rational) Adaptive choice of shifts involves nonlinear term $BB^\top$

♣ Proper Orthogonal Decomposition (functional based)

♣ Balanced Truncation

Back to the reduced Riccati equation

$$T_k^\top \mathbf{Y} + \mathbf{Y} T_k - \mathbf{Y} B_k B_k^\top \mathbf{Y} + C_k^\top C_k = 0 \qquad (*)$$

THEOREM. Let the pair (T_k, B_k) be stabilizable and (C_k, T_k) observable. Then there is a unique solution $\mathbf{Y}_k \geq 0$ of $(*)$ that for each $\widehat{x}_0$ gives the feedback optimal control

$$\widehat{u}_*(t) = -B_k^* \mathbf{Y}_k \exp((T_k - B_k B_k^* \mathbf{Y}_k)t)\widehat{x}_0, \quad t \geq 0$$

for the reduced system.

♣ If there exists a matrix K such that $A - BK$ is dissipative, then the pair (T_k, B_k) is stabilizable.

The dissipating feedback matrix problem

Given

$$\begin{cases} \dot{x} &= Ax - Bu \\ u &= Kx, \end{cases} \tag{1}$$

and A **not** dissipative, find, if it exists, a dissipating feedback matrix K such that the *closed-loop* linear system $\dot{x} = (A - BK)x$ is dissipative.

(Guglielmi, Simoncini, tr 2018)

This means "the field of values of $A - BK$ is all in $\mathbb{C}^-$", that is

$$(A - BK) + (A - BK)^\top < 0$$

Known existence results and parameterization

A classical result (tailored to our setting):

see, e.g., Skelton, Iwasaki & Grigoriadis 1998

THEOREM. Assume B is full column rank. Then

(i) There exists a matrix K satisfying $A + A^\top - BK - (BK)^\top < 0$ if and only if

$$B^\perp (A + A^\top)(B^\perp)^\top < 0 \quad \text{or} \quad BB^\top > 0;$$

(ii) The following parameterization holds

$$K = -R^{-1}B^\top + R^{-\frac{1}{2}} L \Phi^{-\frac{1}{2}},$$

where $L \in \mathbb{R}^{q\times n}$ is an arbitrary matrix such that $\|L\| < 1$ and $R \in \mathbb{R}^{q\times q}$ is an arbitrary positive definite matrix such that $\Phi := (BR^{-1}B^\top - (A + A^\top))^{-1} > 0$.

A counter-example

This parameterization does not seem to include all possible Ks:

EXAMPLE. Consider $Q := A + A^\top = \mathrm{diag}(\alpha, -\alpha)$, with $\alpha > 0$, and $B = e_1 = [1; 0]$. Let us take $R^{-1} = \widehat{\alpha}$ with $\widehat{\alpha} > \alpha$. Then

$$\begin{aligned} \Phi &= (BR^{-1}B^* - Q)^{-1} = \mathrm{diag}(\frac{1}{\widehat{\alpha} - \alpha}, \frac{1}{\alpha}) > 0, \\ \widetilde{B} &= \Phi^{\frac{1}{2}} B R^{-\frac{1}{2}} = \frac{\sqrt{\widehat{\alpha}}}{\sqrt{\widehat{\alpha} - \alpha}} e_1 \end{aligned}$$

with $\|\widetilde{B}\| = \frac{\sqrt{\widehat{\alpha}}}{\sqrt{\widehat{\alpha}-\alpha}} > 1$ for all choices of $\alpha > 0$ and $\widehat{\alpha} > \alpha$. By taking $L = \frac{1}{2}\widetilde{B}$, α and $\widehat{\alpha}$ can be selected so that $\|L\| \geq 1$, while for this choice of L we still have $BK + K^\top B^\top + Q < 0$. $\square$

Thinking again the existence result

$$\mathcal{M} = \begin{bmatrix} (A + A^\top) & B \\ B^\top & 0 \end{bmatrix}$$

- If the matrix $(A + A^\top)$ is negative definite on the kernel of $B^\top$, then $\mathcal{M}$ has exactly q positive and n negative eigenvalues
- The matrix $A + A^\top$ is negative definite on the kernel of $B^\top$ if and only if there exists a $K \in \mathbb{R}^{q \times n}$ such that $W(A - BK) \subset \mathbb{C}^-$

Constructive derivation:

$$\mathcal{M} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} X \\ Y \end{bmatrix} \Lambda, \qquad \Lambda < 0$$

Then

$$K = Y X^{-1} \qquad\qquad (X \text{ nonsingular})$$

Thinking again the existence result. Generalization.

The set of all Ks can be enlarged:

> THEOREM. There exists a matrix K such that $W(A - BK) \subset \mathbb{C}^-$ if and only if the pencil $(\mathcal{M}, \mathcal{D})$ admits n negative eigenvalues for some symmetric and positive definite matrix $\mathcal{D} \in \mathbb{R}^{(n+q)\times(n+q)}$.

Hence, for any $\mathcal{D}$ symmetric and positive definite such that

$$\mathcal{M} \begin{bmatrix} X \\ Y \end{bmatrix} = \mathcal{D} \begin{bmatrix} X \\ Y \end{bmatrix} \Lambda, \qquad \Lambda < 0$$

with $\begin{bmatrix} X \\ Y \end{bmatrix} \in \mathbb{R}^{(n+q)\times n}$ $\mathcal{D}$-orthogonal, we define $K := YX^{-1}$

- Other parameterizations are possible

Computing a (weakly) dissipating feedback of minimal norm

Let $\mathbb{W}^{q\times n}(A,B)$ be the set of dissipating matrices for the pair (A,B)

The problem: *Find* $K \in \mathbb{W}^{q\times n}(A,B)$ *such that*

$$\min_{K\in\mathbb{W}^{q\times n}(A,B)} \|K\|_\star$$

($\star$ = F-norm, 2-norm)

♣ For $K \in \mathbb{W}^{q\times n}(A,B)$, the matrix $A + A^\top - BK - (BK)^\top$ has a zero eigenvalue with multiplicity m, with $0 < m \leq q$

The Linear Matrix Inequality (LMI) optimization problem

♣ LMI framework for the 2-norm:

$$\min_{K \in \mathbb{R}^{q\times n}} \|K\|_2 \qquad \text{subject to}$$

$$A + A^\top - BK - K^\top B^\top \leq 0, \qquad \begin{bmatrix} \gamma I_q & K \\ K^\top & \gamma I_n \end{bmatrix} \geq 0$$

(where $\gamma > 0$ is such that $\|K\|_2 \leq \gamma$)

♣ LMI framework for the F-norm:

$$\min_{K \in \mathbb{R}^{q\times n}} \|K\|_F \qquad \text{subject to}$$

$$A + A^\top - BK - K^\top B^\top \leq 0, \qquad \begin{bmatrix} I & \mathrm{vec}(K) \\ \mathrm{vec}(K)^\top & \gamma \end{bmatrix} \geq 0$$

($\mathrm{vec}(K)$ stacks all columns of K one after the other, so that $\|K\|_F^2 \leq \gamma$)

A simple example

Method	description
GL(m)	2-step functional method with m eigs (Guglielmi-Lubich, '17)
LMI	Matlab basic function for the LMI problem (`mincx`)
Yalmip1	Matlab version of Yalmip with SeDuMi solver (2-norm)
Yalmip2	Matlab version of Yalmip with SeDuMi solver (F-norm)
Pencil	minimization problem with pencil $(\mathcal{M}, \mathcal{D})$

$$A = \begin{bmatrix} -0.2 & 1.6 & 0.2 & 2.6 & -0.4 \\ -0.2 & -0.8 & -1.2 & -0.7 & -1.8 \\ 1.4 & 0.7 & -1.1 & 0.2 & 0.8 \\ 0.3 & 0.8 & 0.1 & -0.1 & -0.9 \\ 0.2 & -0.2 & 0.7 & -1.9 & 0.1 \end{bmatrix}, \qquad B = \begin{bmatrix} 0.6 & 0.5 \\ -0.2 & 0.3 \\ 0.5 & 0 \\ 0.2 & 0.6 \\ 0.6 & -0.6 \end{bmatrix}.$$

$$\lambda_i(\frac{1}{2}(A + A^\top)) = \{-2.4752,\ -1.8301,\ -0.7238,\ 0.6506,\ 2.2785\}$$

A simple example

♣ Numerically optimal dissipating matrices:

$$K_{GL} = \begin{bmatrix} 0.3690 & -0.12149 & 0.34503 & 0.1119 & 0.35065 \\ 1.0340 & 0.66501 & -0.01895 & 1.3640 & -1.2432 \end{bmatrix}$$

and

$$K_{Yalmip2} = \begin{bmatrix} 0.3684 & -0.11954 & 0.35079 & 0.1097 & 0.3467 \\ 1.0118 & 0.65736 & -0.03002 & 1.3995 & -1.2240 \end{bmatrix}$$

♣ Eigs of $S(K) = A + A^\top - BK - (BK)^\top$:

$$\lambda_i(S(K_{GL})) \in \{-2.4765,\ -1.8306,\ -0.72468,\ -2.4 \cdot 10^{-9},\ -1.3 \cdot 10^{-8}\}$$

and

$$\lambda_i(S(K_{Yalmip2})) \in \{-2.4743,\ -1.8298,\ -0.72353,\ -2.4 \cdot 10^{-10},\ 5.0 \cdot 10^{-10}\}$$

A simple example

Comparison of the different methods:

Method	Minimization	$\|K_*\|_2$	$\|K_*\|_F$
GL(2)	F-norm	**2.2166**	**2.3063**
LMI	2-norm	**2.2166**	2.6714
Yalmip1	2-norm	**2.2166**	2.5765
Yalmip2	F-norm	**2.2166**	**2.3063**
Pencil	F-norm	2.2560	2.7585

Note: on harder problems Yalmip always gives smallest minimum

Outlook

♠ Reduced Differential Riccati equations
(see, e.g., Koskela & Mena, tr 2017-2018, Güldogan etal tr 2017)

$$\dot{X}(t) = A^\top X(t) + X(t)A - X(t)BB^\top X(t) + C^T C$$

(work in progress, with G. Kirsten)

♠ Parameterized Algebraic Riccati equations
(see, e.g., Schmidt & Haasdonk, 2018)

♠ Feedback control for nonlinear PDEs by state dependent Riccati equation

$$\begin{aligned} \dot{x}(t) &= f(x(t)) + Bu(t), \qquad f(x) = A(x)x \\ z(t) &= Cx(t) \end{aligned}$$

(work in progress, with A. Alla and D. Kalise)