Stability via reverse I-projections, a symplectic perspective on the computation

Wouter Jongeneel*

June 16, 2023, date of first upload: June 18, 2022

Abstract

This note presents a numerical study of the reverse I-projection as recently proposed to enforce stability in the context of linear system identification. Using classical symplectic machinery we derive a competitive numerical algorithm to compute this projection.

Keywords—algebraic Riccati equation, asymptotic stability, QZ algorithm, symplectic group, system identification
AMS Subject Classification (2020)—22E70, 49N10, 60F10, 65F15, 65F45

1 Introduction

"*While the basic theory is established, research concerning the design and analysis of algorithms for solving Riccati equations is still very active and intense, due to the strong demand from a growing number of applications.*" —Bini, Iannazzo, and Meini [BIM11].

In this note we consider a computational problem arising in linear system identification. In particular, in [JSK23] the authors address the following problem. Let us be given a stochastic discrete-time linear time-invariant system of the form

$$x_{t+1} = \theta x_t + w_t, \quad x_0 \sim \nu, \tag{1.1}$$

where $x_t \in \mathbb{R}^n$ and $w_t \in \mathbb{R}^n$ denote the state and noise at time $t \in \mathbb{N}$, respectively. In addition, θ represents a *fixed* yet *unknown* system matrix, and ν stands for the marginal distribution of the initial state x_0. Moreover, assume that θ is ***asymptotically stable***, *i.e.*, $\theta \in \Theta = \{\theta \in \mathbb{R}^{n\times n} : \rho(\theta) < 1\}$, where $\rho(\theta)$ is the spectral radius.

Now, the goal is to identify θ from a single-trajectory of data $\{\widehat{x}_t\}_{t=0}^{T}$ generated by (1.1). The most common method is to use the least squares estimator

$$\widehat{\theta}_T = \left(\textstyle\sum_{t=1}^{T} \widehat{x}_t \widehat{x}_{t-1}^{\mathsf{T}}\right)\left(\textstyle\sum_{t=1}^{T} \widehat{x}_{t-1} \widehat{x}_{t-1}^{\mathsf{T}}\right)^{-1}. \tag{1.2}$$

Although θ is asymptotically stable, $\widehat{\theta}_T$ is not necessarily an element of Θ, *e.g.*, due to noise or insufficient data. This potential qualitative mismatch is unfortunate from a practical point of view [VODM96, pp. 53-60, 125–129] and theoretically non-trivial to handle due to Θ being non-convex in general.

Overcoming this hurdle led to a large body of work on the statistics of $\widehat{\theta}_T$, projecting $\widehat{\theta}_T$ onto Θ, the representation of stable matrices and more, *e.g.*, see [Mac95; LB02; LB03; BGS08; Van+00; Van+01; Tur+13; ONV13; Ume+18; Sim+18; GKS19; JP19; SR19; JP20; SRD20; NP20; CGS20]. However, none of these works provides a principled and tractable method to obtain an estimator of θ that is guaranteed to be asymptotically stable, with statistical guarantees. We refer the reader to [JSK23] for more comments on related work.

*(Technical report, comments are welcome) The author is with the Risk Analytics and Optimization Chair, École Polytechnique Fédérale de Lausanne (EPFL) and is supported by the Swiss National Science Foundation under the NCCR *Automation*, grant agreement 51NF40_180545. Contact: `wouter.jongeneel@epfl.ch`, `wjongeneel.nl`

Contribution To address the aforementioned gap in the literature, Jongeneel, Sutter, and Kuhn [JSK23] propose a projection method inspired by the theory of large deviations. Their method comes with statistical guarantees, corresponds to solving a single Linear Quadratic Regulator (LQR) problem and yields an estimator guaranteed to be stable. Although no numerical problems are reported, at first sight, the LQR problem at hand seems ill-conditioned. In this note we elaborate on this particular LQR problem and clarify how the structure of the underlying LQR problem in [JSK23] is — and can be further — exploited. It turns out that by employing machinery due to, Moler and Stewart [MS73], Pappas, Laub, and Sandell [PLS80] and Mehl et al. [Meh+09] this structural observation allows for a fast and numerically stable reformulation of their projection.

Structure We start by introducing the reverse I-projection proposed in [JSK23]. Then, we highlight the relation between Riccati equations and a particular generalized eigenvalue problem. The next sections further exploit the closely-related symplectic structure and show how to reformulate the LQR problem in a numerically well-conditioned manner. The note is concluded with comments on the implementation and a variety of numerical experiments.

Notation Given a real matrix $A \in \mathbb{R}^{n\times m}$, A^{T} denotes its transpose, whereas for a complex matrix $Z \in \mathbb{C}^{n\times m}$, Z^{H} denotes its conjugate (Hermitian) transpose. The real matrix inner product $\mathsf{tr}(A^{\mathsf{T}}B)$ is denoted by $\langle A, B\rangle$, the operator norm and spectral radius of matrix A are denoted by $\|A\|_2$ and $\rho(A)$, respectively. The additive group of n-dimensional symmetric matrices is denoted by $\mathsf{Sym}(n)$ and the cone of n-dimensional symmetric positive definite matrices is denoted by $\mathcal{S}^n_{\succ 0}$. The dimension of 0 is either explicitly indicated, *i.e.*, $0_{m\times n} \in \mathbb{R}^{m\times n}$, irrelevant, or implied by the context. Given a matrix $A \in \mathbb{C}^{n\times n}$ with simple eigenvalue $\lambda \in \mathrm{spec}(A)$, left- and right eigenvectors will be normalized and are denoted by w^{H} and v, respectively, *i.e.*, $w^{\mathsf{H}}A = \lambda w^{\mathsf{H}}$, $Av = \lambda v$. All random objects are defined on a measurable space $(\Omega, \mathcal{F})$ equipped with a probability measure $\mathbb{P}_\theta$ parametric in (the fixed) θ. Similarly, the expectation operator with respect to $\mathbb{P}_\theta$ is denoted by $\mathbb{E}_\theta[\cdot]$. We will use *h.o.t.* as an acronym for *higher order terms.*

2 The reverse I-projection

In this section we largely follow [JSK23]. In particular, besides $\theta \in \Theta$, we assume the following throughout.

Assumption 2.1 (Disturbance statistics [JSK23, Assumption II.1])**.** *The disturbances $\{w_t\}_{t\in\mathbb{N}}$ are independent and identically distributed (i.i.d.) and independent of x_0 under $\mathbb{P}_\theta$. The marginal noise distributions are unbiased ($\mathbb{E}_\theta[w_t] = 0$), non-degenerate ($S_w = \mathbb{E}_\theta[w_t w_t^{\mathsf{T}}] \succ 0$ is finite) and have an everywhere positive probability density function.*

To abstract time away, let $\Theta' = \mathbb{R}^{n\times n}$ be the space of realizations of $\widehat{\theta}_T$. Then, inspired by the *nearest stable matrix problem*

$$\Pi_\Theta(\theta') \in \arg\min_{\theta\in\mathsf{cl}\,\Theta} \|\theta' - \theta\|_2^2, \tag{2.1}$$

and the theory of large deviations [Hol08; DZ09], consider the so-called "*rate function*" $I : \Theta'\times\Theta \to [0,\infty]$

$$I(\theta', \theta) = \tfrac{1}{2}\mathsf{tr}\left(S_w^{-1}(\theta' - \theta)S_\theta(\theta' - \theta)^{\mathsf{T}}\right) \tag{2.2}$$

and correspondingly the reverse I-projection

$$\mathcal{P}(\theta') \in \arg\min_{\theta\in\Theta} I(\theta', \theta), \tag{2.3}$$

for some $\theta' \in \Theta'$. Let $\mathsf{dlqr}(A,B,Q,R)$ denote any[1] standard infinite-horizon discrete-time LQR routine that outputs the optimal — with respect to the cost matrices Q and R — feedback gain K such that $\rho(A+BK)<1$, *cf.* [Ber05; Ber07]. It turns out that (2.3) has the following attractive properties.

Theorem 2.2 (The reverse I-projection [JSK23, Theorem II.3]). *Suppose that Assumption 2.1 holds, that the noise is light-tailed as well as stationary and that $\widehat{\theta}_T$ is the least squares estimator (1.2). Then, for any $\theta \in \Theta$ the reverse I-projection defined in (2.3) displays the following properties.*

(i) ***Asymptotic consistency.***

$$\lim_{T\to\infty} \mathcal{P}(\widehat{\theta}_T) = \theta \quad \mathbb{P}_\theta\text{-}a.s.$$

(ii) ***Finite sample guarantee.*** *There are constants $\tau \geq 0$ and $\rho \in (0,1)$ that depend only on θ such that*

$$\mathbb{P}_\theta \left(\|\theta - \mathcal{P}(\widehat{\theta}_T)\|_2 \leq \kappa(S_w) \frac{2\varepsilon n^{\frac{1}{2}}\tau}{\sqrt{1-\rho^2}} \right) \geq 1-\beta$$

for all $\beta, \varepsilon \in (0,1)$ and $T \geq \kappa(S_w)\widetilde{O}(n)\log(1/\beta)/\varepsilon^2$.

(iii) ***Efficient computation.*** *For any $\theta' \notin \Theta$ and $S_w, Q \succ 0$ there is a $p \geq 1$, such that for all $\delta > 0$ we have that*

$$\theta^\star_\delta = \theta' + \mathsf{dlqr}(\theta', I_n, Q, (2\delta S_w)^{-1})$$

is stable and satisfies $\|\mathcal{P}(\theta') - \theta^\star_\delta\|_2 \leq O(\delta^p)$.

2.1 Computing the reverse I-projection

The LQR-based computation (approximately) of $\mathcal{P}(\theta')$ as proposed by Theorem 2.2, *i.e.* $\theta^\star_\delta$, is algebraically characterized by the symmetric positive definite solution P_δ to the ***algebraic Riccati equation***

$$P_\delta = Q + \theta'^{\mathsf{T}} P_\delta \left(I_n + 2\delta S_w P_\delta\right)^{-1} \theta'. \tag{2.4}$$

A solution P_δ to (2.4) immediately translates to $\theta^\star_\delta = (I_n + 2\delta S_w P_\delta)^{-1}\theta'$. Clearly, one *cannot* simply set $\delta = 0$ when $\theta' \notin \Theta$. See [LR95] for a complete account on algebraic Riccati equations, see also [BLW91] for more intimately related topics.

We like to point out that in [JSK23] $Q \succ 0$ is considered such that usual observability conditions trivially hold, *cf.* [Ber05, Chapter 4]. Besides well-known relaxtions to $Q \succeq 0$ based on *detectability* arguments we highlight one more situation. When θ' has no unimodular eigenvalues, $Q = 0_{n\times n}$ suffices to find a stabilizing solution of the corresponding Riccati equation, however, only when $\theta' \in \Theta$, then, this solution is stabilizing *and* still relates to a LQR problem. The observation underlying these statements is that for $Q = 0_{n\times n}$ the algebraic Riccati equation can have multiple solutions, $P_\delta = 0_{n\times n}$ being one of them.

Equation (2.4) provides an algebraic perspective, but in the end, the problem under consideration is the computation of

$$\theta^\star_\delta = \theta' + \mathsf{dlqr}(\theta', I_n, Q, (2\delta S_w)^{-1}) \tag{2.5}$$

for some $Q \succ 0$ and some sufficiently small $\delta > 0$. Both of these parameters are up to the user, for example, in [JSK23] $Q = I_n$ and $\delta = 10^{-9}$ are used successfully. In what follows, as is done below and throughout, to aid the reader, certain potential choices of Q are highlighted by means of a bullet $\bullet$ $(\cdots)$.

[1] For example, consider the MATLAB function `https://ch.mathworks.com/help/control/ref/dlqr.html`, or the (a) Julia Riccati solver `https://github.com/andreasvarga/MatrixEquations.jl/blob/master/src/riccati.jl` that functions as the backend of the Julia $\mathsf{dlqr}(\cdot)$ method. Note, we do not consider cross-terms in the cost.

- (Naïve): $Q = I_n$.

To see why Q is an important parameter, assume for a moment that $2S_w = I_n$, let Q be of the form δQ for some $\delta > 0$ and let P_δ solve the corresponding Riccati equation (2.4). Now see the following

$$P_\delta = \delta Q + \theta'^{\mathsf{T}} P_\delta \left(I_n + \delta P_\delta\right)^{-1} \theta' \iff \tfrac{1}{\delta} P_\delta = Q + \tfrac{1}{\delta}\theta'^{\mathsf{T}} P_\delta \left(I_n + \delta^2 \tfrac{1}{\delta} P_\delta\right)^{-1} \theta'. \tag{2.6}$$

The construction from above shows that scaling Q by δ is equivalent to the original problem under δ^2 instead of δ, *i.e.*, we see that

$$\mathsf{dlqr}(\theta', I_n, \delta Q, \delta^{-1} I_n) = \mathsf{dlqr}(\theta', I_n, Q, \delta^{-2} I_n).$$

Evidently, the former formulation is numerically preferred and motivates some form of scaling

- (Re-scaling): $Q \leftarrow \delta^d Q$, $d \in \mathbb{N}_{\geq 1}$.

Evidently, terms of the form $O(\delta^{-1})$ for arbitrarily small $\delta > 0$ are numerically challenging. As we highlight below, most of the standard $\mathsf{dlqr}(\cdot)$ routines can in fact handle this situation satisfactory. Perhaps more interesting, as "*sufficiently small*" is not well-defined, one might want to input a monotonically decreasing sequence $\{\delta_k\}_k$ and terminate when some appropriate stopping condition is satisfied. To that end, the selection of Q, perhaps as a function of δ, is ought to play a role, *i.e.*, to make sure the computation is numerically stable.

Computational contribution (*informal*). Using a classical symplectic perspective, the dependency on δ^{-1}, *e.g.*, see (2.5), can be reformulated to a problem where δ appears linearly instead. Employing such a framework, this note provides a simple iterative QZ routine — almost as simple as the standard $\mathsf{dlqr}(\cdot)$ routine — to solve for $\theta^\star_\delta \approx \operatorname{argmin}_{\theta \in \Theta} I(\theta', \theta)$ in a fast and numerically stable way. To do so, the symplectic perspective is notably employed towards selecting $Q \in \mathscr{Q}$, from some appropriate function class $\mathscr{Q}$, that aids the numerical computation. We theoretically motivate why $Q(\delta) = 2\delta\theta'^{\mathsf{T}} S_w \theta'$ is appropriate and corroborate this with numerical experiments.

In [JSK23], some computational details are hidden in the proof of [JSK23, Proposition III.6]. This relates to the special case of $\theta' \in \partial\Theta$, but more generally to θ' having unimodular eigenvalues. In that case some nice properties of $I(\theta', \theta)$ deteriorate, in particular, the computation is necessarily an approximation. The next example highlights this and illustrates how δ might appear.

Example 2.3 (Closed-form solutions)**.** *To shed some more light on the computation, consider the scalar problem of evaluating* $\mathsf{dlqr}(1, 1, 1, \delta^{-1})$ *for some* $\delta > 0$. *The corresponding Riccati equation* (2.4) *becomes* $\delta p^2 - \delta p - 1 = 0$. *The positive solution is given by* $p_\delta = (\delta + \sqrt{\delta(\delta + 4)})/(2\delta)$. *Then, as* $\theta^\star_\delta = (1 + \delta p_\delta)^{-1}\theta'$ *we obtain*

$$\theta^\star_\delta = \frac{1}{1 + \delta/2 + \sqrt{\delta(\delta + 4)}/2}.$$

Hence, we observe stability and the preservation of the orientation, i.e., $\operatorname{sgn}\det(\theta^\star_\delta) = \operatorname{sgn}\det(\theta')$, *as given by [JK21; JSK23]. However, we also see that* $\lim_{\delta \downarrow 0} \theta^\star_\delta \in \partial\Theta$. *The reason being that* $I(1, \theta) = \frac{1}{2}(1 - \theta)^2/(1 - \theta^2)$ *does not blow-up for* $\theta \uparrow 1$ *cf [JSK23, Proposition III.6.3)]. Now we compute* $\mathsf{dlqr}(1, 1, q, \delta^{-1})$ *for some* $\delta > 0$ *and* $q > 0$ *and get* $\theta^\star_\delta = 1/(1 + \delta q/2 + \sqrt{\delta(\delta q + 4)}/2)$. *Hence, for arbitrarily small* $q > 0$ *one has* $\theta^\star_\delta = O(1/(1 + \sqrt{\delta}))$. *Recall that the machine precision* $\mu > 0$ *is defined as the smallest rational number such that* $1 + \mu > 1$ *on that particular machine. Hence, we must have* $\delta > \mu^2$ *to assert asymptotic stability of* $\theta^\star_\delta$, *numerically. At last, let* $q = 0$ *and* $|\theta'| \neq 1$. *Now we find that the stabilizing solution to* (2.4) *becomes*

$$p_\delta = \begin{cases} 0 & \textit{if } \theta'^2 < 1 \\ (\theta'^2 - 1)/\delta & \textit{if } \theta'^2 > 1 \end{cases},$$

i.e., $p_\delta = 0$ *does not result in stabilization of* θ' *with* $\theta'^2 > 1$*. Summarizing, for* $\theta'^2 < 1$ *one has* $\lim_{\delta\downarrow 0}\theta^\star_\delta = \theta'$*, as it should be, and for* $\theta'^2 > 1$ *one has* $\lim_{\delta\downarrow 0}\theta^\star_\delta = 1/\theta'$*, which displays the symmetry as envisioned in [JSK23, Figure 2].*

3 A generalized eigenvalue problem

Example 2.3 hints at the non-trivial dependency of $\theta^\star_\delta$ on δ and in particular that a closed-form solution becomes quickly prohibitive.

Pappas, Laub, and Sandell [PLS80] show that any solution to an (discrete-time) algebraic Riccati equation like (2.4) can be represented directly via a solution to a generalized eigenvalue problem. This allows for some additional algebraic insights beyond dynamic programming formulations and eventually for an efficient and stable algorithm. To keep the work remotely self-contained we briefly show how a generalized eigenvalue problem arises.

To start, define the pair of matrices $S_1, S_2 \in \mathbb{R}^{2n\times 2n}$ by

$$S = \{S_1, S_2\} = \left\{ \begin{pmatrix} \theta' & 0_{n\times n} \\ -Q & I_n \end{pmatrix}, \begin{pmatrix} I_n & 2\delta S_w \\ 0_{n\times n} & \theta'^{\mathsf{T}} \end{pmatrix} \right\}. \tag{3.1}$$

If useful, the dependency on δ is made explicit, *e.g.*, $S_2(\delta)$. Then, the ***generalized eigenvalues*** of the ***matrix pencil*** $S_1 - \lambda S_2$ are defined as the set $\operatorname{spec}(S_1, S_2) = \{\lambda \in \mathbb{C} : \det(S_1 - \lambda S_2) = 0\}$, generalizing the notation $\operatorname{spec}(A)$ for the spectrum of a matrix A[2]. Now, consider the generalized eigenvalue problem

$$S_1 x = \lambda S_2 x, \tag{3.2}$$

for $x \in \mathbb{C}^{2n}$ a generalized eigenvector with corresponding eigenvalue $\lambda \in \mathbb{C}$. By looking at the pencil

$$S_1 - \lambda S_2 = \begin{pmatrix} \theta' - \lambda I_n & -2\delta\lambda S_w \\ -Q & I_n - \lambda\theta'^{\mathsf{T}} \end{pmatrix},$$

we see that $\lambda = 0$ is a solution to (3.2) if and only if $0 \in \operatorname{spec}(\theta')$. Moreover, it can be shown, *e.g.*, see [PLS80, Theorem 4], that the generalized eigenvalues satisfying (3.2) come in reciprocal pairs. As the next lemma will show, the motivation for the study of (3.2) follows from the fact that a subset of these generalized eigenvectors can be directly mapped to P_δ satisfying (2.4).

Towards this construction, we will write a solution to (3.2) using a formulation reminiscent of the standard Jordan Normal form, specifically, define $X^s, X^u \in \mathbb{C}^{2n\times n}$, $X_{ij} \in \mathbb{C}^{n\times n}$ for $i, j = 1, 2$ and $J^s, J^u \in \mathbb{C}^{n\times n}$ by $S_1 X = S_2 X J$, that is

$$\begin{pmatrix} \theta' & 0_{n\times n} \\ -Q & I_n \end{pmatrix}\begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix} = \begin{pmatrix} I_n & 2\delta S_w \\ 0_{n\times n} & \theta'^{\mathsf{T}} \end{pmatrix}\begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix}\begin{pmatrix} J^s & 0_{n\times n} \\ 0_{n\times n} & J^u \end{pmatrix} \tag{3.3}$$

for $X^s = [X_{11}^{\mathsf{H}}\ X_{21}^{\mathsf{H}}]^{\mathsf{H}}$, $X^u = [X_{12}^{\mathsf{H}}\ X_{22}^{\mathsf{H}}]^{\mathsf{H}}$. If there is no λ solving (3.2), with $|\lambda| = 1$, we know that since the eigenvalues come in reciprocal pairs, there is one strictly (asymptotically) *stable* n-dimensional eigenspace $\operatorname{im}(X^s)$ and one strictly *unstable* n-dimensional eigenspace $\operatorname{im}(X^u)$. Throughout, we will reserve the block J^s to denote the Jordan Normal form corresponding to the stable eigenspace $\mathcal{X}^s = \operatorname{im}(X^s) \subseteq \mathbb{C}^{2n}$. Now we can highlight a link between (2.4) and a solution to (3.2).

Lemma 3.1 (Structure of solutions to (2.4) [PLS80, Lemma 1])**.**

(i) All solutions of (2.4) *are of the form* $X_a X_b^{-1}$*, where* $[X_a^{\mathsf{H}}\ X_b^{\mathsf{H}}]^{\mathsf{H}} \in \mathbb{C}^{2n\times n}$ *compromises a set of* n *generalized eigenvectors corresponding to* (3.2).

[2] It is also common to consider instead of λ a pair (α, β) such that $\det(\beta S_1 - \alpha S_2) = 0$ and map it back to λ via α/β. This is employed in algorithms.

(ii) Let P_δ be a solution to (2.4) and let $X \in \mathbb{C}^{2n\times 2n}$ be the set of generalized eigenvectors corresponding to (3.2), where $X^s = [X_{11}^{\mathsf{H}}\ X_{21}^{\mathsf{H}}]^{\mathsf{H}} \in \mathbb{C}^{2n\times n}$ denotes a basis for the stable eigenspace. Then, $P_\delta = X_{21}X_{11}^{-1}$ and $\theta_\delta^\star = X_{11}J^s X_{11}^{-1} \in \Theta$.

From now on, when we talk about the pair (X_{11}, X_{21}), this is in the sense of (3.3) and Lemma 3.1. At first, Lemma 3.1 does not seem to bring a lot of new information to the table. However, we will see in the upcoming sections that the representation $P_\delta = X_{21}X_{11}^{-1}$, and the matrix pencil formulation in general, allows for numerical and analytical insights beyond what we could get using a common dynamic programming formulation.

3.1 Numerical stability Given the results from Lemma 3.1, one might consider computing P_δ using a generalized eigenvalue solver. However, since the computation of eigenvectors is numerically unstable (see Example 5.1 below), this approach is not preferred. In fact, constructing the Jordan normal form is not a continuous matrix decomposition. Towards a more stable algorithm we should consider a unitary basis. Again, to be somewhat self-contained we discuss the merits via an example.

Example 3.2 (A numerically stable basis [GL13, p. 354]). *Consider for some $0 < \varepsilon \ll 1$ the matrix*

$$A = \begin{pmatrix} 1+\varepsilon & 1 \\ 0 & 1-\varepsilon \end{pmatrix}. \tag{3.4}$$

We can diagonalize A and obtain $A = T\Lambda T^{-1}$ for $\Lambda = \operatorname{diag}(1+\varepsilon, 1-\varepsilon)$ and

$$T = \begin{pmatrix} 1 & 1 \\ -2\varepsilon & 0 \end{pmatrix}, \quad \kappa_2(T) = \frac{1+2\varepsilon^2+(\varepsilon^2+1)^{1/2}}{1+2\varepsilon^2-(\varepsilon^2+1)^{1/2}}.$$

Then from [GL13, p. 100] we know that a floating-point computation (finite arithmetic), denoted $\mathsf{fl}(\cdot)$, of a similarity transformation yields $\mathsf{fl}(T^{-1}AT) = T^{-1}AT + E$ with $\|E\|_2 \lesssim \mu\kappa_2(T)\|A\|_2$, for μ being (proportional to) machine precision. Indeed, for $\varepsilon \to 0$ the error bound deteriorates since the condition number $\kappa_2(T)$ explodes. If instead of $T \in \mathsf{GL}(n,\mathbb{R})$ we could use a $Q \in \mathsf{O}(n,\mathbb{R})$, then, we minimize $\|E\|_2$ since $\kappa_2(Z) \geq 1\ \forall Z \in \mathbb{R}^{n\times n}$ while $\kappa_2(Q) = 1\ \forall Q \in \mathsf{O}(n,\mathbb{R})$.

Fortunately, one can always find a unitary basis, to be self-contained we sketch a proof.

Lemma 3.3 (Real Schur decomposition [GL13, Theorem 7.4.1]). *For any $A \in \mathbb{R}^{n\times n}$, there exists a $Q \in \mathsf{O}(n,\mathbb{R})$ such that*

$$Q^{\mathsf{T}}AQ = R = \begin{pmatrix} R_{11} & R_{12} & \cdots & R_{1m} \\ 0 & R_{22} & \cdots & R_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & R_{mm} \end{pmatrix} \tag{3.5}$$

where all R_{ii} blocks are either scalars containing the eigenvalues of A or 2×2 matrices containing complex conjugate eigenvalue pairs of A.

Proof (sketch). Let A have some eigenvalue $\lambda \in \operatorname{spec}(A)$ with eigenspace $E_\lambda \subseteq \mathbb{R}^n$. As the matrix A acts linearly on $\mathbb{R}^n$, we can construct the direct sum $\mathbb{R}^n = E_\lambda \oplus E_\lambda^\perp$ and find some orthonormal basis matrices Q_1 and Q_2 for E_λ and $E_\lambda^\perp$, respectively. By construction we have

$$\begin{pmatrix} Q_1 & Q_2 \end{pmatrix}^{\mathsf{T}} A \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} = \begin{pmatrix} J_\lambda & Q_1^{\mathsf{T}}AQ_2 \\ 0 & Q_2^{\mathsf{T}}AQ_2 \end{pmatrix},$$

where J_λ is a Jordan block of appropriate size. Now repeat the procedure and decompose $Q_2^{\mathsf{T}}AQ_2$. See that the Schur decomposition is not unique when $\dim(E_\lambda) > 1$ for some $\lambda \in \operatorname{spec}(A)$. □

Note, due to possible 2×2 blocks on the diagonal R could fail to be upper triangular in the general sense. As such, $R \in \mathbb{R}^{n \times n}$ in (3.5) is called upper *quasi*-triangular.

Example 3.4 (Example 3.2 continued)**.** *The matrix A in (3.6) is already in the real Schur form, however, now consider*

$$A = \begin{pmatrix} 1+\varepsilon & 1 \\ \varepsilon & 1-\varepsilon \end{pmatrix}, \tag{3.6}$$

one can show that the corresponding matrix R is given by

$$R = \begin{pmatrix} 1+(\varepsilon(\varepsilon+1))^{1/2} & 1-\varepsilon \\ 0 & 1-(\varepsilon(\varepsilon+1))^{1/2} \end{pmatrix}. \tag{3.7}$$

Next, Lemma 3.3 is extended to the generalized eigenvalue setting.

Lemma 3.5 (Generalized real Schur decomposition [GL13, Theorem 7.7.2])**.** *For any $A, B \in \mathbb{R}^{n \times n}$ there exist $Q, Z \in \mathsf{O}(n, \mathbb{R})$ such that $Q^\mathsf{T} AZ$ is upper quasi-triangular and $Q^\mathsf{T} BZ$ is upper triangular.*

The reason of interest in Lemma 3.5 is that the generalized spectra $\operatorname{spec}(A, B)$ and $\operatorname{spec}(Q^\mathsf{T} AZ, Q^\mathsf{T} BZ)$ are equal, *cf.* (3.2).

Now, classically, we could write down a solution to a generalized eigenvalue problem using the Jordan Normal form as in Lemma 3.1, *e.g.*, $AX = BXJ$. The following celebrated result, as initiated by [Lau79], tells us that we can also use the Schur decomposition, which as highlighted before, has superior numerical properties.

Lemma 3.6 ([PLS80, Theorem 8a])**.** *Consider for the pair (S_1, S_2) as given by (3.1) its generalized real Schur decomposition as proposed in Lemma 3.5, i.e., there are matrices $Q, Z \in \mathsf{O}(2n, \mathbb{R})$ such that $Q^\mathsf{T} S_1 Z$ is upper quasi-triangular and $Q^\mathsf{T} S_2 Z$ is upper triangular. Then, all solutions of (2.4) are of the form $P = U_{21} U_{11}^{-1}$, for U being defined as*

$$U = \begin{pmatrix} U_{11} \\ U_{21} \end{pmatrix} = Z \begin{pmatrix} I_n \\ 0_{n \times n} \end{pmatrix} = \begin{pmatrix} Z_{11} \\ Z_{21} \end{pmatrix}. \tag{3.8}$$

We refer to Lemma 3.6 as the ***QZ method***, or QZ algorithm (to compute P_δ). The original QZ algorithm, that is, to compute the generalized real Schur decomposition is due to Moler and Stewart [MS73]. To provide intuition, a sketch of the proof of Lemma 3.6 is given below.

Proof (sketch). Since $\lambda(A, B)$ and $\lambda(Q^\mathsf{T} AZ, Q^\mathsf{T} BZ)$ are equal, we can instead of $S_1 X^s = S_2 X^s J^s$ look at $S_1 U = S_2 U R$, with $R \in \mathbb{R}^{n \times n}$ being stable and upper quasi-triangular[3] and $U \in \mathbb{R}^{2n \times n}$ being defined by (3.8). See [PLS80, p. 637] for more on the construction. Then, from Lemma 3.1 we know that P is of the form $X_{21} X_{11}^{-1}$. Since R is stable, construct its Jordan Normal form, that is $R = TJT^{-1}$. From there we obtain $S_1 UT = S_2 UTJ$. Indeed, after an application of Lemma 3.1 we get that $P = U_{21} T (U_{11} T)^{-1} = U_{21} U_{11}^{-1}$. Using this observation we can extend Lemma 3.1 to any type of basis, *e.g.*, see early remarks in [Fat69; Wil71]. □

From the proof of Lemma 3.6 we also see that $\theta_\delta^\star = (U_{11} T) J (U_{11} T)^{-1} = U_{11} R U_{11}^{-1}$. So, if desired, one could skip the computation of P_δ. To elaborate on the possibility of using *any* basis related to the stable eigenspace $\mathcal{X}$, see that for any invertible matrix T we have

$$S_1 \begin{pmatrix} X^s & X^u \end{pmatrix} \begin{pmatrix} T & 0 \\ 0 & I_n \end{pmatrix} = S_2 \begin{pmatrix} X^s & X^u \end{pmatrix} \begin{pmatrix} T & 0 \\ 0 & I_n \end{pmatrix} \begin{pmatrix} T & 0 \\ 0 & I_n \end{pmatrix}^{-1} \begin{pmatrix} J^s & 0 \\ 0 & J^u \end{pmatrix} \begin{pmatrix} T & 0 \\ 0 & I_n \end{pmatrix}$$

such that if the basis X^s for $\mathcal{X}^s$ follows the transformation $X^s T$, we still have

$$\begin{aligned} (X_{21} T)(X_{11} T)^{-1} = X_{21} X_{11}^{-1} &= P_\delta, \\ (X_{11} T)(T^{-1} J^s T)(X_{11} T)^{-1} = X_{11} J^s X_{11}^{-1} &= \theta_\delta^\star. \end{aligned} \tag{3.9}$$

[3] Here one exploits group properties of upper-triangular matrices.

Note, (3.9) also displays a potential numerical complication not immediately resolved by the QZ method; one needs still solve a linear system.

4 The symplectic group

Although, practically, we want a method capable of computing $\theta_\delta^\star$ for any $\theta' \in \Theta'$, we will see that the analysis of the pair $(\theta_\delta^\star, P_\delta)$ simplifies once we move to a slightly less general setting. Specifically, assume for the moment that $\theta' \in \mathsf{GL}(n,\mathbb{R}) = \{\theta \in \mathbb{R}^{n\times n} : \det(\theta) \neq 0\}$. Note, as θ' corresponds to a realization of the least squares estimator $\widehat{\theta}_T$ and $\mathsf{GL}(n,\mathbb{R})$ is dense in $\mathbb{R}^{n\times n}$, such an assumption is not overly restrictive. Moreover, the Q matrices we derive using this analysis can be implemented without θ' being invertible.

Next, define $\Omega \in \mathbb{R}^{2n\times 2n}$ by[4]

$$\Omega = \begin{pmatrix} 0_{n\times n} & -I_n \\ I_n & 0_{n\times n} \end{pmatrix}.$$

Here, Ω defines an anti-symmetric billinear form $\omega : \mathbb{C}^{2n} \times \mathbb{C}^{2n} \to \mathbb{C}$ via $\omega(x,y) = \langle x, \Omega y\rangle$, *e.g.*, $\omega(x,x) = 0$. Then, define the real ***symplectic group*** by

$$\mathsf{Sp}(2n,\mathbb{R}) = \{M \in \mathbb{R}^{2n\times 2n} : M^{\mathsf{T}}\Omega M = \Omega\}.$$

The corresponding Lie algebra is given by $\mathfrak{sp}(2n,\mathbb{R}) = \{X \in \mathbb{R}^{2n\times 2n} : X^{\mathsf{T}}\Omega + \Omega X = 0\}$, with the standard matrix (commutator) bracket, *i.e.* $[A,B] = AB - BA$ for appropriately sized matrices A and B. Elements X of $\mathfrak{sp}(2n,\mathbb{R})$ satisfy $\Omega X \in \mathsf{Sym}(2n)$ and are called ***Hamiltonian matrices***. A useful property of the symplectic and Hamiltonian matrices is that for any $M \in \mathsf{Sp}(2n,\mathbb{R})$ and $X \in \mathfrak{sp}(2n,\mathbb{R})$ we have that $M^{-1}XM \in \mathfrak{sp}(2n,\mathbb{R})$. When $M \in \mathsf{Sp}(2n,\mathbb{R})$ we speak of M being Ω-symplectic. Moreover, we speak of a subspace $\mathcal{Y}$ being M-invariant, when $M\mathcal{Y} \subseteq \mathcal{Y}$. Now, since $\theta' \in \mathsf{GL}(n,\mathbb{R})$ it follows from (3.1) that $S_2 \in \mathsf{GL}(2n,\mathbb{R})$ such that $S_2^{-1}S_1$ is well-defined and in fact $S_2^{-1}S_1 \in \mathsf{Sp}(2n,\mathbb{R})$. Specifically, we can define the curve $M : \mathbb{R} \to \mathsf{Sp}(2n,\mathbb{R})$ by

$$\delta \mapsto M(\delta) = S_2^{-1}S_1 = \begin{pmatrix} \theta' + 2\delta S_w \theta'^{-\mathsf{T}} Q & -2\delta S_w \theta'^{-\mathsf{T}} \\ -\theta'^{-\mathsf{T}} Q & \theta'^{-\mathsf{T}} \end{pmatrix}. \tag{4.1}$$

Using this notation, see that for $\delta_1, \delta_2 \in \mathbb{R}$ we have by the group properties of $\mathsf{Sp}(2n,\mathbb{R})$ that

$$M^{\mathsf{T}}(\delta_2)M^{\mathsf{T}}(\delta_1)\Omega M(\delta_1)M(\delta_2) = M^{\mathsf{T}}(\delta_2)\Omega M(\delta_2) = \Omega.$$

At times we will write M_δ instead of $M(\delta)$ to simplify notation. Similarly, when δ is a curve itself, $\delta(t)$ might be written as δ_t. One could interpret δ as if it is a *structure-preserving* perturbation. Differently put, we call (S_1, S_2) a ***symplectic pair*** (or symplectic pencil), *i.e.*, $S_1\Omega S_1^{\mathsf{T}} = S_2\Omega S_2^{\mathsf{T}}$. To continue, inspired by the Schur complement, $M(\delta)$ can be decomposed as follows:

$$M(\delta) = \begin{pmatrix} I_n & -2\delta S_w \\ 0_{n\times n} & I_n \end{pmatrix} \begin{pmatrix} \theta' & 0_{n\times n} \\ 0_{n\times n} & \theta'^{-\mathsf{T}} \end{pmatrix} \begin{pmatrix} I_n & 0_{n\times n} \\ -Q & I_n \end{pmatrix}, \tag{4.2}$$

e.g., see also [Meh88, Proposition 2.36]. Indeed, whilst repressing notational dependence on δ, then from (4.2) we immediately see that $\theta' \in \mathsf{GL}(n,\mathbb{R}) \implies \det(M) = 1$. The fact that the eigenvalues of M come in reciprocal pairs can be seen from $\Omega M \Omega^{-1} = M^{-\mathsf{T}}$. Moreover, when $Q = 2\delta S_w$ see that (4.2) depicts a *symplectic similarity transformation*, denoted $M(\delta) = T(\delta)^{\mathsf{T}} D(\theta) T(\delta)$, $T(\delta), D(\theta) \in \mathsf{Sp}(2n,\mathbb{R})$. This observation hints at a seemingly natural choice of Q.

- (Symplectic symmetry): $Q = 2\delta S_w$.

[4] Instead of Ω one frequently uses the notation J for the symplectic matrix. However, we use J already to denote Jordan blocks.

It is imperative to remark that symplectic matrices are inherently hard to handle numerically. Reasons being, either one likes to have some eigenvalues on the unit disk $\mathbb{D}_1$, *e.g.*, in mechanics [Arn10, Chapter 8], or by reciprocity of its eigenvalues, M is likely to have unstable eigenvalues, even eigenvalues "*at infinity*". As highlighted in the previous section, we are only concerned with a stable subspace and by construction unimodular eigenvalues are ruled out, this eliminates some of the aforementioned numerical problems.

Ideally, one understands how the selection (perturbation) of the pair (δ, Q) influences the spectrum of M. Although work in that direction exists [MBO97; SM16; SMM20], explicit dependencies are non-trivial and to that end we consider a different line of attack in the upcoming sections.

5 Ω-Lagrangian subspaces

We know from Lemma 3.6 that *any* basis for $\mathcal{X}^s = \mathrm{im}(X^s)$ will do to construct the pair $(\theta^\star_\delta, P_\delta)$. It is of great interest to understand how this subspace $\mathcal{X}^s \subseteq \mathbb{C}^{2n}$ will change as we perturb δ, *i.e.*, how we will evolve over a Grassmannian manifold. As promoted in the previous section, understanding the effect of δ aids in appropriately selecting Q as a function of δ. Unfortunately, perturbation theory with respect to the standard eigenvector basis is limited *cf.* [Lax07, Theorem 8, p. 130], [Kat95, Chapter 2]).

Example 5.1 (Discontinuous eigenvector basis)**.** *Consider the real* 2×2 *matrix*

$$A = \begin{pmatrix} \alpha & 1 \\ 0 & \beta \end{pmatrix}$$

for some $\alpha, \beta \in \mathbb{R}$*. Now see that the corresponding Jordan decomposition is discontinuous (and ill-conditioned) in the pair* (α, β)*:*

$$A = \begin{cases} \begin{pmatrix} 1 & 1 \\ 0 & \beta-\alpha \end{pmatrix}\begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix}\begin{pmatrix} 1 & (\alpha-\beta)^{-1} \\ 0 & (\beta-\alpha)^{-1} \end{pmatrix} & \alpha \neq \beta \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \alpha & 1 \\ 0 & \beta \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} & \alpha = \beta \end{cases}.$$

From Example 5.1 it appears that even selecting Q to be simply linear in δ is not necessarily numerically stable. However, Example 5.1 is concerned with general matrices, it turns out that the situation improves when we focus on symplectic matrices.

To overcome the difficulties with studying eigenvectors we will use tools which were initially brought to life within the realms of classical mechanics, *e.g.*, see [AM08, Section 5.3]. In particular, we use the results due to Mehl et al. [Meh+09].

A subspace $W \subset \mathbb{C}^{2n}$ is called ***Ω-Lagrangian*** when $\dim(W) = n$ and $\omega(x,y) = 0\ \forall x, y \in W$ [LR95, p. 274]. Equivalently, a subspace $W \subseteq \mathbb{C}^{2n}$ is Ω-Lagrangian when for the orthogonal complement of W with respect to ω, that is $W^\perp = \{x \in \mathbb{C}^{2n} : \omega(x,y) = 0\ \forall y \in W\}$, we have $W = W^\perp$ [Lee13, p. 566].

Lemma 5.2 (Qualitative properties of $\mathcal{X}^s$)**.** *The subspace* $\mathcal{X}^s \subseteq \mathbb{C}^{2n}$ *is* Ω*-Lagrangian and* $M(\delta)$*-invariant for all* $\delta > 0$*.*

Proof. From Lemma 3.1, specifically (3.3), it follows that

$$\theta' = X_{11} J^s X_{11}^{-1} + 2\delta S_w X_{21} J^s X_{11}^{-1}, \tag{5.1a}$$

$$X_{21} X_{11}^{-1} = \theta'^{\mathsf{T}} X_{21} J^s X_{11}^{-1} + Q. \tag{5.1b}$$

Combining (5.1a) and (5.1b) yields

$$X_{21}X_{11}^{-1} = (X_{11}J^sX_{11}^{-1})^{\mathsf{T}} X_{21}X_{11}^{-1}(X_{11}J^sX_{11}^{-1}) + Q + 2\delta(X_{21}J^sX_{11}^{-1})^{\mathsf{T}} S_w (X_{21}J^sX_{11}^{-1}).$$

Hence, $X_{21}X_{11}^{-1}$ is symmetric such that

$$\operatorname{im}\begin{pmatrix} I_n \\ X_{21}X_{11}^{-1} \end{pmatrix}$$

is Ω-Lagrangian since for any $x, y \in \mathcal{X}^s$:

$$\omega(x,y) = u^{\mathsf{H}} \begin{pmatrix} I_n & (X_{21}X_{11}^{-1})^{\mathsf{T}} \end{pmatrix} \begin{pmatrix} 0_{n\times n} & -I_n \\ I_n & 0_{n\times n} \end{pmatrix} \begin{pmatrix} I_n \\ X_{21}X_{11}^{-1} \end{pmatrix} v = 0 \quad \forall u, v \in \mathbb{C}^n.$$

Now, by multiplying both u and v from the left with $X_{11} \in \mathsf{GL}(n,\mathbb{C})$ we see that $\mathcal{X}^s$ is Ω-Lagrangian as well. Moreover, by the Jordan normal form construction of X^s it follows that $\mathcal{X}^s$ is $M(\delta)$-invariant. □

As in [Meh+09], to study the stability of these invariant Lagrangian subspaces define the ***gap*** between subspaces $W \subseteq \mathbb{C}^m$ and $U \subseteq \mathbb{C}^m$ by

$$\operatorname{gap}(W,U) = \|P_W - P_U\|_2, \tag{5.2}$$

for P_W being the orthogonal projection operator, mapping any $x \in \mathbb{C}^m$ onto W. The next lemma tells us that if the gap (5.2) between two subspaces of equal dimension can be made arbitrarily small, then, there are generators (basis matrices) which are arbitrary close in norm.

Lemma 5.3 (Continuity)**.** *Let $W, W_\varepsilon \subseteq \mathbb{C}^m$ be p-dimensional subspaces generated by $W = \operatorname{im}(X)$ and $W_\varepsilon = \operatorname{im}(X_\varepsilon)$ for a fixed full-rank $X \in \mathbb{C}^{m\times p}$ with $p \leq m$ and some full-rank $X_\varepsilon \in \mathbb{C}^{m\times p}$, such that $\operatorname{gap}(W,W_\varepsilon) < \varepsilon$ for $\varepsilon > 0$. Then, if for some sequence $\{W_\varepsilon\}_\varepsilon$ we have $\lim_{\varepsilon\to 0}\operatorname{gap}(W,W_\varepsilon) = 0$ there exists a sequence $\{X_\varepsilon\}_\varepsilon$ such that $W_\varepsilon = \operatorname{im}(X_\varepsilon)$ and $\lim_{\varepsilon\to 0}\|X - X_\varepsilon\|_2 = 0$.*

Proof. First, we write (5.2) explicitly as:

$$\operatorname{gap}(W,W_\varepsilon) = \left\|X(X^{\mathsf{H}}X)^{-1}X^{\mathsf{H}} - X_\varepsilon(X_\varepsilon^{\mathsf{H}}X_\varepsilon)^{-1}X_\varepsilon^{\mathsf{H}}\right\|_2 < \varepsilon. \tag{5.3}$$

Observe that the evaluation of (5.3) for X_ε or $X_\varepsilon T$, $T \in \mathsf{GL}(p,\mathbb{C})$ is equivalent. Then, since the projection map $\Pi : \mathbb{C}^{m\times p} \to \mathbb{C}^{m\times m}$ defined by $\Pi(X) = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}$ is continuous at full-rank matrices, which compromise an open set of $\mathbb{C}^{m\times p}$, the result follows. □

Regarding stability we use the following definition from [Meh+09], for the moment we overload the meaning of δ.

Definition 5.4 (Ω-stable invariant subspace [Meh+09, Definition 3.1])**.** *Let $M \in \mathbb{R}^{2n\times 2n}$ be Ω-symplectic and $\mathcal{X} \subseteq \mathbb{C}^{2n}$ be a M-invariant Ω-Lagrangian subspace. Then, $\mathcal{X}$ is Ω-stable if $\forall\, \varepsilon > 0$ there is a $\delta > 0$ such that if $M' \in \mathbb{R}^{2n\times 2n}$ is Ω-symplectic and $\|M - M'\|_2 < \delta$, then, there is a M'-invariant Ω-Lagrangian subspace $\mathcal{X}' \subseteq \mathbb{C}^m$ such that $\operatorname{gap}(\mathcal{X},\mathcal{X}') < \varepsilon$.*

To continue, let $\mathbb{D}_1 = \{z \in \mathbb{C} : |z| < 1\}$ be a subset of the complex plane $\mathbb{C}$ and recall that ***unimodular eigenvalues*** are elements of $\partial\mathbb{D}_1$. Now we state one of the main results from [Meh+09]:

Lemma 5.5 ([Meh+09, Corollary 7.3])**.** *Let $M \in \mathbb{R}^{2n\times 2n}$ be Ω-symplectic. Then, if and only if M has no unimodular eigenvalues there exist unique Ω-stable M-invariant Ω-Lagrangian subspaces $\mathcal{X}^s, \mathcal{X}^u \subseteq \mathbb{C}^{2n}$ such that $\lambda(M|_{\mathcal{X}^s}) \subseteq \mathbb{D}_1$ and $\lambda(M|_{\mathcal{X}^u}) \subseteq \mathbb{C} \setminus \operatorname{cl}(\mathbb{D}_1)$.*

The ramification of Lemma 5.5 is as follows. By construction, $M(\delta)$ does not have unimodular eigenvalues, otherwise $M(\delta)$ would not lead to a stabilizing solution of the Riccati equation (2.4). As such, a combination of Lemma 5.2 and Lemma 5.5 shows that $\mathcal{X}^s$ is stable — in the sense of Definition 5.4 — with respect to feasible perturbations in the pair (δ, Q). Hence, discontinuities as in Example 5.1 do not pertain to our setting and we find that Q should be at least continuous in δ, *e.g.*, $Q = 2\delta S_w$ indeed.

- (Ω-stability): $Q(\delta)$ being continuous in δ.

The next section shows the benefits of selecting Q to be differentiable in δ.

6 Symplectic perturbation theory

"*We use neither the differentiable structure of the symplectic group nor the Lie algebra structure of Hamiltonian matrices.*" —[BIM11].

In standard LQR parlance, we only consider perturbations in the pair of cost matrices (Q, R), in our notation of Section 2, $(Q, (2\delta S_w)^{-1})$. Most of the related work on Riccati perturbations present *implicit* bounds, see for example [Sun98]. By exploiting the Lie group structure of $\mathsf{Sp}(2n, \mathbb{R})$ we get more *explicit* insights. To do so we take a continuous-time approach.

6.1 Perturbations and the exponential map The Lie group point of view is viable since the exponential map $\exp : \mathfrak{g} \to \mathsf{G}$ reduces to the *matrix* exponential when G is a *matrix* Lie group, *e.g.*, for any $t \in \mathbb{R}$ and $X \in \mathfrak{g}$ one has

$$\exp(tX) = \sum_{k\geq 0} \frac{1}{k!}(tX)^k,$$

see [Var84; DK99]. To that end, assume that $\exp(t_1 X)M(\delta(t_0)) = M(\delta(t_1))$ holds for some $X \in \mathfrak{sp}(2n, \mathbb{R})$ implicitly defined by $M(\delta(t))$, a path in $\mathsf{Sp}(2n, \mathbb{R})$ parametrized by $t \in [t_0, t_1]$ for a sufficiently small interval. In what follows we consider without loss of generality the non-empty interval $[0, s]$. We also assume for simplicity of the exposition that $\delta(t)$ is affine in t, *e.g.*, $\delta(t) = \delta_0 - t$, for $t \in [0, \delta_0)$.

Going beyond mere continuity, as derived to be necessary in the previous section, now also assume that Q is a smooth function (at least differentiable) in δ, *i.e.*, $Q(\delta) \succ 0$ for all $\delta > 0$ and $dQ : \mathbb{R}_{>0} \to \mathsf{Sym}(n)$. Under this assumption, compute

$$\partial_\delta M(\delta)|_{\delta=\delta'} = \begin{pmatrix} 2S_w\theta'^{-\mathsf{T}}Q(\delta') + 2\delta' S_w \theta'^{-\mathsf{T}}\, \partial_\delta Q(\delta)|_{\delta=\delta'} & -2S_w\theta'^{-\mathsf{T}} \\ -\theta'^{-\mathsf{T}}\, \partial_\delta Q(\delta)|_{\delta=\delta'} & 0_{n\times n} \end{pmatrix}. \tag{6.1}$$

Then, since $d_t M(\delta(t))|_{t=0} = XM(\delta(0))$ we find by computing $X = \partial_t \delta(t)|_{t=0}\, \partial_\delta M(\delta)|_{\delta(0)}\, M(\delta(0))^{-1}$ that

$$X = \partial_t\delta(t)|_{t=0} \begin{pmatrix} 2\delta_0 S_w \theta'^{-\mathsf{T}}\, \partial_\delta Q(\delta)|_{\delta=\delta_0}\, \theta'^{-1} & 4\delta_0^2 S_w\theta'^{-\mathsf{T}}\, \partial_\delta Q(\delta)|_{\delta=\delta_0}\, \theta'^{-1}S_w - 2S_w \\ -\theta'^{-\mathsf{T}}\, \partial_\delta Q(\delta)|_{\delta=\delta_0}\, \theta'^{-1} & -2\delta_0\theta'^{-\mathsf{T}}\, \partial_\delta Q(\delta)|_{\delta=\delta_0}\, \theta'^{-1}S_w \end{pmatrix}.$$

As $\partial_\delta Q(\delta) \in \mathsf{Sym}(n)$, one readily verifies that $\Omega X \in \mathsf{Sym}(2n)$ and $X \in \mathfrak{sp}(2n, \mathbb{R})$ indeed. As $M(\delta(t))$ is fixed, and $\exp : \mathfrak{sp}(2n, \mathbb{R}) \to \mathsf{Sp}(2n, \mathbb{R})$ is not surjective, this verification is necessary. We see that if Q is constant, X is nilpotent. Moreover, ignoring $\partial_t\delta(t)|_{t=0}$ by our linearity assumption, we can conveniently factor[5] X as follows

$$X = \begin{pmatrix} 2\delta_0 S_w \\ -I_n \end{pmatrix} \theta'^{-\mathsf{T}}\, \partial_\delta Q(\delta)|_{\delta=\delta_0}\, \theta'^{-1} \begin{pmatrix} I_n & 2\delta_0 S_w \end{pmatrix} + \begin{pmatrix} -I_n \\ 0_{n\times n} \end{pmatrix} \begin{pmatrix} 0_{n\times n} & 2S_w \end{pmatrix}. \tag{6.2}$$

Let us write (6.2) as $X = A_1 + A_2$ for $A_1, A_2 \in \mathfrak{sp}(2n, \mathbb{R})$. Now observe that both $A_1^2 = 0$ and $A_2^2 = 0$, that is, A_1 and A_2 are nilpotent. However, $(A_1A_2)^k$ is non-zero for any integer $k \geq 0$, *i.e.*, $[A_1, A_2] \neq 0$. Hence, the dependency of Q on δ, even just linearly, makes $\exp(tX)$ non-trivial. In particular, we find that

$$[A_1, A_2] = \begin{pmatrix} -2S_w\theta'^{-\mathsf{T}}\, \partial_\delta Q(\delta)|_{\delta=\delta_0}\, \theta'^{-1} & 0_{n\times n} \\ 0_{n\times n} & \theta'^{-\mathsf{T}}\, \partial_\delta Q(\delta)|_{\delta=\delta_0}\, \theta'^{-1}2S_w \end{pmatrix}. \tag{6.3}$$

One could interpret (6.3) as $[X, A_2] = [A_1 + A_2, A_2] = [A_1, A_2]$, which in its turn, one can interpret in a Lie bracket context *cf.* [Arn10, Chapter 8], [NS90, Chapter 3]. With this in mind, the

[5] With respect to the factorization of X, note that $\begin{pmatrix} 4\delta & 8\delta^2 \\ -2 & -4\delta \end{pmatrix} = \begin{pmatrix} 2\sqrt{2}\delta \\ -\sqrt{2} \end{pmatrix} \begin{pmatrix} \sqrt{2} & 2\sqrt{2}\delta \end{pmatrix}$.

assignment of $Q(\delta) = d^{-1}\delta^{d}\theta'^{\mathsf{T}}(2S_w)^{-1}\theta'$, for some integer $d \geq 2$, results in a "*well-conditioned*" bracket (6.3), regardless of the pair (θ', S_w), with (arbitrarily) small norm,

$$[A_1, A_2] = \begin{pmatrix} -\delta_0^{d-1} I_n & 0_{n\times n} \\ 0_{n\times n} & \delta_0^{d-1} I_n \end{pmatrix}, \tag{6.4}$$

i.e., for $\delta_0 < 1$ one damps out higher-order terms that could otherwise result from perturbations in δ.

- (Higher-order damping): $Q(\delta) = d^{-1}\delta^{d}\theta'^{\mathsf{T}}(2S_w)^{-1}\theta'$, for some integer $d \geq 2$.

The downside of such a selection of $Q(\delta)$ is that one might damp-out beneficial parts as well.

Under this selection of $Q(\delta)$ we get

$$X = \partial_t \delta(t)|_{t=0} \begin{pmatrix} \delta_0^{d} I_n & 2\delta_0^{d+1} S_w - 2S_w \\ -\delta_0^{d-1}(2S_w)^{-1} & -\delta_0^{d} I_n \end{pmatrix}. \tag{6.5}$$

Regarding the spectrum of X, we see that $\sum_i \lambda_i(X) = 0$ indeed. Moreover, while ignoring $\partial_t \delta(t)|_{t=0}$ again, evaluating[6] $\det(X - \lambda I_{2n}) = 0$ immediately implies that all eigenvalues of X are $\pm(\delta_0^{d-1})^{1/2}$. We see that the contribution of S_w is rendered out, for the better or worse.

6.1.1 Zassenhaus' formula To keep the note remotely self-contained, we make the previous statement regarding higher-order terms more explicit. First, observe that for any $A_1, A_2 \in \mathfrak{sp}(2n, \mathbb{R})$ one can expand the exponential map as follows

$$\exp(A_1 + A_2) = I_{2n} + A_1 + A_2 + \tfrac{1}{2}A_1^2 + \tfrac{1}{2}A_2^2 + \tfrac{1}{2}A_1A_2 + \tfrac{1}{2}A_2A_1 + \text{h.o.t.}$$

Now see that

$$\begin{aligned} &\exp(A_1)\exp(A_2)\exp(-\tfrac{1}{2}[A_1, A_2]) = \\ &\quad (I_{2n} + A_1 + \tfrac{1}{2}A_1^2 + \text{h.o.t.})(I_{2n} + A_2 + \tfrac{1}{2}A_2^2 + \text{h.o.t.})(I_{2n} - \tfrac{1}{2}[A_1, A_2] + \text{h.o.t.}) \\ &\qquad = I_{2n} + A_1 + A_2 + \tfrac{1}{2}A_1^2 + \tfrac{1}{2}A_2^2 + \tfrac{1}{2}A_1A_2 + \tfrac{1}{2}A_2A_1 + \text{h.o.t.}. \end{aligned}$$

This simple observation illustrates *Zassenhaus' formula* as formally stated below.

More generally, let $X, Y \in \mathfrak{g}$ for some Lie algebra $\mathfrak{g}$, then the solution $Z(X,Y) \in \mathfrak{g}$ to

$$\exp(X)\exp(Y) = \exp(Z(X,Y)), \tag{6.6a}$$

assuming it exists, is given by a series of the form

$$Z(X,Y) = X + Y + \tfrac{1}{2}[X,Y] + \tfrac{1}{12}\left([[X,Y],Y] + [[Y,X],X]\right) + \cdots, \tag{6.6b}$$

e.g., see [Var84, § 2.15]. This formula is called the ***Baker-Campbell-Hausdorff formula***. In particular, let $\|\cdot\|_{\mathfrak{g}}$ be a sub-multiplicative norm on $\mathfrak{g}$, then, for any $X, Y \in \mathfrak{g}$ such that $\|X\|_{\mathfrak{g}} + \|Y\|_{\mathfrak{g}} < \log(2)$ the series (6.6b) converges to a solution (6.6a) [BBM20, Proposition 2.2]. For example, consider the *operator* $\|\cdot\|_2$ or *Frobenius* norm $\|\cdot\|_F$, respectively, on $\mathfrak{sp}(2n, \mathbb{R})$.

It turns out that a more useful expansion exists, bearing Zassenhaus' name. Let $\mathfrak{g}(A_1, A_2) \subseteq \mathfrak{sp}(2n, \mathbb{R})$ be the free Lie (sub)-algebra generated by $A_1, A_2 \in \mathfrak{sp}(2n, \mathbb{R})$. Then, there exists the unique decomposition

$$\exp(A_1 + A_2) = \exp(A_1)\exp(A_2)\prod_{n=2}^{\infty}\exp(C_n(A_1, A_2)), \tag{6.7}$$

[6] Due to the simple structure of (6.5), $\det(X - \lambda I_{2n}) = \det((\lambda - \delta_0^{d})(\delta_0^{d} + \lambda)I_n - (\delta_0^{d-1} - \delta_0^{2d})I_n)$.

where $C_n \in \mathfrak{g}(A_1, A_2)$ is a homogeneous Lie polynomial of degree n[7]. See for example [CMN12] for more background information on ***Zassenhaus' formula*** (6.7). Evidently, when $[A_1, A_2] \approx 0$, then (6.7) implies that $\exp(A_1+A_2) \approx \exp(A_1)\exp(A_2)$. As mentioned before, although convenient from an analysis point of view, *i.e.*, higher-order perturbation are largely independent of the pair (θ', S_w), these annihilated higher-order terms could have been potentially beneficial. The next section takes a different point of view.

6.2 Approximate geodesics In this section we employ standard machinery from Riemannian geometry, *e.g.*, see [Arn10, Appendix 2], to find Q as a function of δ such that $M(\delta(t))$ is "*approximately*" a geodesic on $\mathsf{Sp}(2n, \mathbb{R})$. We are motivated to do so by existing multiplicative perturbation results [MBO97; SM16; SMM20].

For any Lie group G, let $L_g : \mathsf{G} \to \mathsf{G}$ be the ***left-translation*** defined by $L_g(h) = gh$ for all $g, h \in \mathsf{G}$. Now, a vector field $X \in \mathfrak{X}(\mathsf{G})$ is ***left-invariant*** when $(L_g)_\star X = X$, *i.e.*, $d(L_g)h(X_h) = X_{gh}$ [Lee13, p. 189]. We will again focus on the real symplectic group $\mathsf{Sp}(2n, \mathbb{R})$, yet not immediately on M as given by (4.1).

Given any $X \in \mathfrak{sp}(2n, \mathbb{R})$, recall that $g\mathfrak{sp}(2n, \mathbb{R}) \simeq T_g\mathsf{Sp}(2n, \mathbb{R})$[8] for all $g \in \mathsf{Sp}(2n, \mathbb{R})$. This allows for defining a left-invariant (Riemannian) metric g[9] on $T_g\mathsf{Sp}(2n, \mathbb{R})$ via

$$\mathsf{g}_g(X, Y) = \langle X, Y\rangle_g = \langle d(L_{g^{-1}})_g(X), d(L_{g^{-1}})_g(Y)\rangle_e \tag{6.8}$$

for all $X, Y \in T_g\mathsf{Sp}(2n, \mathbb{R})$ and any $g \in \mathsf{Sp}(2n, \mathbb{R})$. Here we will simply use the Frobenius-metric, $\langle A, B\rangle = \mathsf{tr}(A^\mathsf{T} B)$, as induced from $\mathbb{R}^{n\times n}$. Similarly, one could define those objects using the right-translation operator $R_g : \mathsf{G} \to \mathsf{G}$. Equipped with the metric g, $(\mathsf{Sp}(2n, \mathbb{R}), \mathsf{g})$ is a Riemannian manifold.

Let $\mathcal{T} \ni t \mapsto M(t) \in \mathsf{Sp}(2n, \mathbb{R})$ be some curve (not necessarily as in (4.1)), in particular, a one-parameter subgroup[10] defined via the exponential map e^{tX}, $X \in \mathfrak{sp}(2n, \mathbb{R})$ that acts on $M(0)$, generally put as $\exp(tX)M(0) = M(t)$ for some reference point $M(0) \in \mathsf{Sp}(2n, \mathbb{R})$. This construction gives rise to the differential equation $\dot{M}(t) = XM(t)$, *e.g.*, a right-invariant vector field on $\mathsf{Sp}(2n, \mathbb{R})$.

Now let $B_i \in \mathfrak{sp}(2n, \mathbb{R})$, $i \in \mathcal{I}$, form a basis for $\mathfrak{sp}(2n, \mathbb{R})$, *i.e.*, $|\mathcal{I}| = \dim(\mathfrak{sp}(2n, \mathbb{R}))$. In the case of $\mathfrak{sp}(2n, \mathbb{R})$, any Hamiltonian matrix $H \in \mathfrak{sp}(2n, \mathbb{R})$ can be written as

$$H = \begin{pmatrix} A & B \\ C & -A^\mathsf{T} \end{pmatrix}, \quad A \in \mathbb{R}^{n\times n},\ B, C \in \mathsf{Sym}(n).$$

Therefore, $\dim(\mathfrak{sp}(2n, \mathbb{R})) \leq 2n^2 + n$. The basis $\{B_i\}_{i\in\mathcal{I}}$ can be used to construct a set of right-invariant vector fields $\{X^{(i)}\}_{i\in\mathcal{I}} \subset \mathfrak{X}(\mathsf{G})$, defined by $X^{(i)}_g = (R_g)_\star B_i$.

We say that a vector field X on a Riemannian manifold (M, g), under the metric g, is a ***Killing vector field*** when $\mathscr{L}_X \mathsf{g} = 0$ [Jos11, Definition 2.3.7], that is, the ***Lie derivative*** of g with respect to X vanishes. Let $[\cdot, \cdot]_{\mathscr{L}}$ denote the Lie bracket, then it follows from [Lee13, Corollary 12.33] that for left-invariant vector fields Y, Z and the metric (6.8) one has

$$\mathscr{L}_X \mathsf{g}(Y, Z) = -\mathsf{g}([X, Y]_{\mathscr{L}}, Z) - \mathsf{g}(Y, [X, Z]_{\mathscr{L}}). \tag{6.9}$$

Now, as right- and left-invariant vector fields on any Lie group commute, one has by our choice of g that $\mathscr{L}_X \mathsf{g} = 0$. Hence, when X is right-invariant, it is a Killing vector field under the metric g.

[7] For example, when $n = 2$, C_2 is a polynomial with basis elements $[A_1, A_1]$, $[A_2, A_2]$, $[A_1, A_2]$ and $[A_2, A_1]$. By the Lie algebra structure, this collapses to a polynomial in $[A_1, A_2]$.

[8] That is, $T_g\mathsf{Sp}(2n, \mathbb{R}) = \{gX : X \in \mathfrak{sp}(2n, \mathbb{R})\} = \{X : X^\mathsf{T}\Omega g + g^\mathsf{T}\Omega X = 0\}$.

[9] Common notation for such a metric would be "g", yet we use g to denote elements of generic Lie groups G.

[10] The map $\varphi(t) = \exp(tX)$ induces a group homomorphism between $(\mathbb{R}, +)$ and $(\mathsf{G}, \cdot)$.

Then, let $\gamma : [0, s] \to \mathsf{Sp}(2n, \mathbb{R})$ be a geodesic. By construction (of the Killing vector fields), we have that $\langle (R_{\gamma(t)})_\star B_i, \dot\gamma(t)\rangle_{\gamma(t)} = C_i(t)$ is conserved over time. This can be seen, for example, from (6.9) by plugging in the connection ∇. The conserved quantity can be rewritten as

$$C_i(t) = \langle B_i\gamma(t), \dot\gamma(t)\rangle_{\gamma(t)} = \langle \gamma(t)^{-1}B_i\gamma(t), \gamma(t)^{-1}\dot\gamma(t)\rangle_e. \tag{6.10}$$

Now consider the curve $t \mapsto \gamma(t) = M(t) = \exp(tX)M(0)$, we aim to argue when a curve of such a form can be a geodesic. Recall that in general we cannot assume that $M(0) = I_{2n}$. Plugging this curve into (6.10) yields

$$\begin{aligned} C_i(t) =& \langle M_0^{-1}e^{-tX}B_ie^{tX}M_0, M_0^{-1}e^{-tX}Xe^{tX}M_0\rangle_e \\ =& \langle M_0^{-1}e^{-tX}B_ie^{tX}M_0, M_0^{-1}XM_0\rangle_e. \end{aligned} \tag{6.11}$$

Then, differentiating with respect to time results in

$$\begin{aligned} 0 = d_tC_i(t) =& \langle M_t^{-1}[B_i, X]M_t, M_0^{-1}XM_0\rangle_e \\ =& \langle M_0^{-1}e^{-tX}[B_i, X]e^{tX}M_0, M_0^{-1}XM_0\rangle_e \\ =& \langle M_0^{-1}e^{-tX}[B_i, X]e^{tX}M_0, M_0^{-1}e^{-tX}Xe^{tX}M_0\rangle_e. \end{aligned}$$

At last, by the trace-inequality (**??**) we have that $\langle [B_i, X], X\rangle_e = 0 \implies d_tC_i(t) = 0$.

Lemma 6.1 (Approximately geodesic)**.** *Let X be as in (6.2), and set $Q(\delta) = 2\delta\theta'^{\mathsf{T}}S_w\theta'$, then, for all $i \in \mathcal{I}$*

$$\langle [B_i, X], X\rangle_e \le O\left(\mathsf{tr}(\delta^4 S_w^8)^{1/2}\right)\mathsf{tr}(B_i). \tag{6.12}$$

Proof. First, see that $[X, X^{\mathsf{T}}] = 0 \implies \langle [B_i, X], X\rangle_e = 0$ for all $i \in \mathcal{I}$. As before, we can omit $\partial_t\delta(t)|_{t=0}$, such that we can conveniently write down $X \in \mathfrak{sp}(2n, \mathbb{R})$ as

$$X = \begin{pmatrix} \delta EF & \delta^2 EFE - E \\ -F & -\delta FE \end{pmatrix}, \quad E = 2S_w,\ F = \theta'^{-\mathsf{T}}\,\partial_\delta Q(\delta)|_{\delta=\delta_0}\,\theta'^{-1}.$$

Note that both E and F are symmetric. Under this compact notation and the selection of Q such that $F = E$, one obtains

$$\begin{aligned} XX^{\mathsf{T}} &= \begin{pmatrix} \delta^2E^4 + E^2 + 2\delta^2E^4 + \delta^4E^6 & -\delta E^3 + \delta E^3 - \delta^3E^5 \\ -\delta E^3 + \delta E^3 - \delta^3E^5 & E^2 + \delta^2E^4 \end{pmatrix} \\ X^{\mathsf{T}}X &= \begin{pmatrix} \delta^2E^4 + E^2 & \delta E^3 - \delta E^3 + \delta^3E^5 \\ \delta E^3 - \delta E^3 + \delta^3E^5 & \delta^2E^4 + E^2 + 2\delta^2E^4 + \delta^4E^6 \end{pmatrix} \end{aligned}$$

and subsequently

$$[X, X^{\mathsf{T}}] = \begin{pmatrix} 2\delta^2E^4 + \delta^4E^6 & -2\delta^3E^5 \\ -2\delta^3E^5 & -2\delta^2E^4 - \delta^4E^6 \end{pmatrix}.$$

In particular, we have $\|[X, X^{\mathsf{T}}]\|_F^2 = 2\mathsf{tr}\left((2\delta^2E^4 + \delta^4E^6)^2 + (2\delta^3E^5)^2\right)$. Such that (6.12) follows since $\langle [B_i, X], X\rangle_e \le \|[X, X^{\mathsf{T}}]\|_F\mathsf{tr}(B_i)$. □

As exp : $\mathsf{Sp}(2n, \mathbb{R}) \to \mathfrak{sp}(2n, \mathbb{R})$ is not surjective, $M(\delta(t))$ might be impossible to reach via the application of $\exp(tX)$ to some $M(\delta(0))$. However, the curve $M(\delta(t))$ can be arbitrarily close to being the image of the exponential map.

Lemma 6.2 (Approximately exponential)**.** *Let $\theta' \in \Theta'$, $S_w \succ 0$, assume that $\rho(S_w) < 1$ and consider $M(\delta(t))$ as in (4.1) for $\delta(t) = \delta_0 - t$, $t \in [0, \delta_0)$ and $\delta(0) = \delta_0 \in (0, 1)$. Set $Q(\delta) = 2\delta\theta' S_w\theta'^{\mathsf{T}}$, then, for all $t \in (0, \delta_0)$*

$$\|M(\delta(t)) - \exp(tX)M(\delta(0))\|_\infty \le \delta_0^4\|S_w^4\|_\infty. \tag{6.13}$$

Proof. One can show that the for the curve $t \mapsto M(\delta(t)) = M(t)$ we have

$$\dot{M}(t) = \lim_{h \downarrow 0} \frac{M(\delta(t+h)) - M(\delta(t))}{h} = XM(t) + \text{h.o.t.},$$

where the higher-order terms are of the order $O((\delta_0 - t)^3 S_w^4)$ (due to $\rho(S_w) < 1$). Now let $\dot{N}(\delta(t)) = XN(\delta(t))$ be a surrogate equation that *does* satisfy the exponential equation, that is, $N(\delta(t)) = \exp(tX)N(\delta(0))$, with $N(\delta(0)) = M(\delta(0))$. As

$$\int_0^t (\delta_0 - s)^3 S_w^4 ds = \frac{1}{4}(\delta_0^4 - (\delta_0 - t)^4) S_w^4 \preceq \delta_0^4 S_w^4,$$

one can easily bound $\|M(\delta(t)) - N(\delta(t))\|_\infty$, resulting immediately in the uniform bound (6.13). □

Lemma 6.1 in combination with Lemma 6.2 indicates that if $\delta_0 \in (0,1)$ and S_w is sufficiently *"small"*, *e.g.*, $\rho(S_w) < 1$, then $Q(\delta) = 2\delta\theta'^{\mathsf{T}} S_w \theta'$ results in $\exp(tX)M(\delta(0))$ being approximately geodesic for sufficiently small t.

- (Approximately geodesic): $Q(\delta) = 2\delta\theta'^{\mathsf{T}} S_w \theta'$.

Conceptually, the geodesic perspective is appealing from a numerical point of view as by the variational properties of such a curve and the first-order perturbation theory [MBO97; SM16; SMM20] one would expect little erratic behaviour.

7 Algorithms and numerical experiments

7.1 Algorithms There are as many algorithms to solve algebraic Riccati equations, as these equations have solutions in general.

Besides the aforementioned *QZ method* — which is frequently used in standard $\mathsf{dlqr}(\cdot)$ routines — we highlight a few other approaches. The most straight-forward method to approximately solve for $\theta_\delta^\star$ is by means of *value iteration* [LR95; Ber05; Ber07], *i.e.*,

$$P_{\delta,k+1} = Q + \theta'^{\mathsf{T}} P_{\delta,k} \left(I_n + 2\delta S_w P_{\delta,k}\right)^{-1} \theta'. \tag{7.1}$$

for some suitable choice of $P_{\delta,0}$, *e.g.*, $P_{\delta,0} = Q$. Speed-ups are possible using *doubling* [LX06]. The so-called *square-root method* due to Lu, Lin, and Pearce [LLP99] proceeds as follows, let $H = (S_1 + S_2)^{-1}(S_1 - S_2)$, then, $\text{im}(H - \text{sqrt}(H^2)) = \mathcal{X}^s$. Although efficient, we clearly face potential problems with conditioning due to $\theta \in \Theta$ and $\delta \downarrow 0$. Another relatively simple method is the *sign method*, see for example [GL86]. Regarding the *QZ method.* Exploiting scaling akin to (2.6) is known to allow for the QZ method to be better conditioned [GKL92], ideally, $\|P_\delta\|_2$ is small [GKL92, Lemma 1], or differently put, $\delta = \|P_\delta\|_2$. To that end, see Lemma 7.1 below. For related benchmark problems see [BLM97] and for an extensive overview of numerical algorithms see [BIM11].

7.1.1 Implementing the QZ algorithm Here, we highlight the computational aspects of the generalized Schur decomposition for a *fixed* $\delta > 0$. The QZ algorithm is employed to construct the decomposition as set forth in Lemma 3.6, the algorithm is backward stable and requires $O(n^3)$ flops in general [GL13, Algorithm 7.7.3]. However, we need to extract n columns from Z corresponding to the *stable* eigenspace, which in general, can be any of the $2n$ columns. Hence, we seek a pair of orthogonal matrices (Q', Z') such that $(Q'^{\mathsf{T}} Q^{\mathsf{T}} S_1 Z Z', Q'^{\mathsf{T}} Q^{\mathsf{T}} S_2 Z Z')$ *does* have the desired order, by somewhat of a convention, the first n columns of ZZ' span $\mathcal{X}^s$. Note, the generalized eigenvalues are easily accessible via the diagonals of $Q^{\mathsf{T}} S_1 Z$ and $Q^{\mathsf{T}} S_2 Z$. Initial stable reordering schemes, which do not make the computational complexity worse, were proposed in [VD81].

Algorithm 1 For a fixed $\delta > 0$, the computation of P_δ, the solution to the algebraic equation (2.4), and the corresponding stabilized system matrix $\theta^\star_\delta$ as given by (2.5) (pseudo-Julia code).

```
Input: $\delta \in \mathbb{R}_{>0}$, $Q \in \mathcal{S}^n_{\succ 0}$, $\theta' \in \mathbb{R}^{n \times n}$, $S_w \in \mathcal{S}^n_{\succ 0}$ and $(S_1(\delta, Q, \theta'), S_2(\delta, \theta', S_w))$ as in (3.1).
Initial step of the QZ algorithm: QZ = schur(S1,S2).
Select the stable eigenvalues: select = abs.(QZ.alpha./QZ.beta).< 1.
Reorder the generalized Schur decomposition: QZreord = ordschur(QZ,select).
Construct $U$ via (3.8) U = QZreord.Z[:,1:n] and $R$ via R = (S2*U)\(S1*U).
Output: $P_\delta = U_{21}U_{11}^{-1}$, $\theta^\star_\delta = U_{11}RU_{11}^{-1}$ and QZreord.
```

See [Kre05, Chapter 2] for more on the QZ algorithm and specifically [Kre05, Section 2.7] for more on numerically stable reordering. To actually perform the computation, we need just a handful of LAPACK [And+99] routines, which are easily accessible via Julia [Bez+17]. Specifically, we can call the schur$(\cdot)$ and ordschur$(\cdot)$ functions in Julia to compute an initial generalized Schur decomposition of (S_1, S_2) and, if desired, reorder the corresponding (generalized) eigenvalues. From this (updated) orthogonal Z, and the definition of U in (3.8), we obtain P_δ by solving $P_\delta U_{11} = U_{21}$. We like to remark that if the (generalized) eigenvalues of the pair (S_1, S_2) are close to $\partial\mathbb{D}_1$ one might want to consider a QZ algorithm that *does* takes the symplectic structure into account [BF98]. By the coercive nature of the rate function (1.2) we effectively avoid this problem *almost surely*.

Regarding scalability, already in 1984, the Schur method, *e.g.*, Algorithm 1, was deemed to be reliable for $n \approx 100$ [AL84]. However, since we generically need $O(n^3)$ flops and $O(n^2)$ memory, for large scale problems one might want to exploit the structure of θ' or resort to different schemes like proposed in [GL91; BF11]. Besides, as indicated in [AL84; Bye87], some problem instances might benefit from a combination of algorithms, *e.g.*, a Schur decomposition and Newton's method, for example to refine a solution. For now we will conclude that an efficient and stable method exists to solve (2.4). Although this observation in itself is not new, it is striking that a non-convex optimization problem, derived from a moderate deviation principle, can be solved using a basic routine from numerical linear algebra.

In Algorithm 1 we present (sub-optimal) pseudo-Julia-code[11] to compute P_δ. One can easily skip the computation of P_δ and directly compute $\theta^\star_\delta$ if desired. Note that when θ' is invertible, R is given by $U^\mathsf{T} M(\delta) U$. Better yet, after re-ordering, $\theta^\star_\delta$ is directly given by the left-most upper-block of $M(\delta)$.

7.1.2 Stopping conditions

Most of the aforementioned indicates that one should use a "*sufficiently small*" $\delta \in (0,1)$ but only this information does not yield a practical algorithm. Given some monotonically decreasing sequence $\{\delta_k\}_k \subset \mathbb{R}_{>0}$ such that $\delta_k \downarrow 0$ for $k \to +\infty$, we propose to use one of the following *absolute* or *relative* stopping conditions

$$|\rho(\theta^\star_{\delta_k}) - \rho(\theta^\star_{\delta_{k-1}})| \leq \epsilon_{\text{abs}}, \quad \frac{|\rho(\theta^\star_{\delta_k}) - \rho(\theta^\star_{\delta_{k-1}})|}{|\delta_{k-1} - \delta_k|} \leq \epsilon_{\text{rel}},$$

respectively, for some $\epsilon_{\text{abs}} > 0$ and $\epsilon_{\text{rel}} > 0$. This approach can be motivated by the fact that the QZ method outputs the spectrum of $\theta^\star_\delta$ (implicitly via α and β) and by the following result together with the fact that $\theta^\star_\delta = (I_n + 2\delta S_w P_\delta)^{-1}\theta'$.

Lemma 7.1 (Ordering of P_δ)**.** *Given some appropriate triple* (θ', Q, S_w)*, let* $\delta_1, \delta_2 \in \mathbb{R}_{>0}$ *be such that* $\delta_1 \geq \delta_2$*, then,* $\delta_1 P_{\delta_1} \succeq \delta_2 P_{\delta_2}$*, for* P_{δ_i} *solving* (2.4) *under* δ_i*.*

Proof. Pre-multiplying (2.4) by δ_i yields

$$\delta_i P_{\delta_i} = \delta_i Q + {\theta'}^\mathsf{T} \delta_i P_{\delta_i} \left(I_n + 2S_w \delta_i P_{\delta_i}\right)^{-1} \theta'.$$

[11] See https://docs.julialang.org/en/v1/stdlib/LinearAlgebra/#LinearAlgebra.Schur

Hence, without loss of generality we consider the problem of solving $\mathsf{dlqr}(\theta', I_n, \delta_i Q, (2S_w)^{-1})$, In fact, by a change of coordinates $x_t = (2S_w)^{1/2}z_t$, with respect to (1.1), we can consider simply $\mathsf{dlqr}(\widetilde{\theta}', I_n, \delta_i Q, I_n)$, for $\widetilde{\theta}' = S_w^{-1/2}\theta' S_w^{1/2}$. Now consider the iterative (dynamic programming) approach to solve for P_{δ_i}, that is, (7.1). Set $P_{\delta_i,0} = \delta_i Q$ such that $P_{\delta_1,0} \succeq P_{\delta_2,0}$. Now, by (7.1) and an inductive argument, this ordering prevails. □

7.2 Iteratively balanced QZ algorithm The motivation for having an algorithm that uses prior knowledge is twofold. First of all, going back to the motivation of this work, the reverse I-projection maps the least squares estimator $\widehat{\theta}_T$ of θ to an asymptotically stable matrix. Assuming that more data is coming, the next estimator, that is $\widehat{\theta}_{T+1}$ is likely to be close to $\widehat{\theta}_T$, so the computation of its reverse I-projection should exploit information of the previous projection step. Moreover, the proposed computation of the reverse I-projection hinges on iteratively computing a generalized Schur decomposition parametric in δ_k, where δ_k is an element of a sequence such that $\delta_k \to 0$. Again, ideally, the decomposition under δ_{k+1} uses information regarding the decomposition under δ_k.

Given a sequence of symplectic pairs $\{A_k, B_k\}_{k\geq 0} \subset \mathbb{R}^{2n\times 2n} \times \mathbb{R}^{2n\times 2n}$, we are tasked with computing the generalized Schur decomposition for each step k. If (A_{k-1}, B_{k-1}) is sufficiently close — which we will make more precise shortly — to (A_k, B_k), then one might hope to reuse knowledge of the previous factorization and thereby speed-up the process. Part of this motivation stems from the backward stability of the QZ algorithm. The "*balancing*" work due to Ward [War81] is one of the earliest in this area. Unfortunately, as also stated in [Kre05, p. 92], balancing can make things worse.

First, consider doing the following, given a pair (A_k, B_k) compute the initial generalized Schur decomposition, which yields the pair (Q_k, Z_k). Next, (if needed) perform a reordering step and update the QZ decomposition, that is, we apply an additional pair of matrices (Q_k', Z_k'). The resulting decomposition $Q_k'^{\mathsf{T}} Q_k^{\mathsf{T}} A_k Z_k Z_k'$, $Q_k'^{\mathsf{T}} Q_k^{\mathsf{T}} B_k Z_k Z_k'$ is used. Then, a new pair (A_{k+1}, B_{k+1}) is received and pre-multiplied by the previous QZ transformation, that is, we start the QZ algorithm from $Q_k'^{\mathsf{T}} Q_k^{\mathsf{T}} A_{k+1} Z_k Z_k'$, $Q_k'^{\mathsf{T}} Q_k^{\mathsf{T}} B_{k+1} Z_k Z_k'$.

Evidently, this can only work if the sequence $\{A_k, B_k\}_{k\geq 0}$ is sufficiently well-behaved. The hope is that this pre-processing step results in (numerical) structure that can be exploited.

Here, we could consider exploiting the algebra formed by upper-triangular matrices. Let $Q_k^{\mathsf{T}} B_k Z_k$ and $Q_k^{\mathsf{T}} B_{k+1} Z_k$ both be upper-triangular. Then, $Q_k^{\mathsf{T}} B_k Z_k Q_k^{\mathsf{T}} B_{k+1} Z_k$ is upper-triangular. In fact, $Q_k^{\mathsf{T}} B_k B_{k+1}^{-1} Q_k$ must be upper-triangular. Differently put, $B_k B_{k+1}^{-1}$ must be *triangulizable* by means on an orthogonal similarity transformation. This is rather a weak condition, yet merely a necessary one in case B_{k+1} is invertible. In fact, $S_2(\delta_k)S_2(\delta_{k+1})^{-1}$ is by construction already upper-triangular, for any choice of δ_k, δ_{k+1}. The non-trivial part is the fact that $S_1(\delta)$ is lower block-triangular, while $S_2(\delta)$ is upper block-triangular.

To see what can go wrong in principle without exploiting prior QZ steps, consider the pair (S_1, S_2) with $n = 1$. Then for S_1, the Schur decomposition yields in general

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} \theta' & 0 \\ q & 1 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & q \\ 0 & \theta' \end{pmatrix}.$$

However, when $q > 0$ is chosen as a function of δ and thereby becomes arbitrarily small for $\delta \downarrow 0$, the Q and Z matrices tend (discontinuously) to I_2. If initialized using our proposed method, this is avoided.

The next example elaborates on this observation.

Example 7.2 (Obstruction to convergence)**.** *Let us compute $\theta^\star_\delta$ by means of the QZ algorithm, e.g., Algorithm 1, but without using prior knowledge. Let $\{\delta_k\}_{k\geq 0} \subset \mathbb{R}_{>0}$ be a sequence converging monotonically to 0. If there is a k such that* $\max\{\|Q(\delta_k)\|_\infty, \|2\delta_k S_w\|_\infty\} = \mu$*, then, we claim that $\theta^\star_{\delta_k}$ does not necessarily converge to something meaningful. See that one can construct the (a)* *generalized Schur decomposition explicitly. For example, let $Q_1^\mathsf{T}\theta' Q_1 = R_1$ be the Schur decomposition of θ' and similarly, let $Q_2^\mathsf{T}\theta'^\mathsf{T} Q_2 = R_2$ be the Schur decomposition of θ'^T. As the off-diagonal (block) terms of the pair (S_1, S_2) are (numerically) zero, a potential output of the QZ algorithm could simply be $Q = \operatorname{diag}(Q_1, Q_2)$, $Z = \operatorname{diag}(Q_1, Q_2)$. Such a solution will be useless in general and can lead to $\rho(\theta^\star_{\delta_k})$ blowing up. The crux is that the QZ algorithm does not lead to a unique decomposition in general.*

Now we highlight how the obstruction from Example 7.2 can be overcome, while also speeding up the process.

In practice one would consider a sequence $\{\delta_k\}_{k\geq 0}$, yet for expositional simplicity, assume that $\delta(t)$ is curve, *e.g.*, $t \mapsto \delta(t) = \delta_0 \exp(-t)$ such that $Q(\delta(t)) \to 0_{n\times n} \in \partial\mathcal{S}^n_{\succ 0}$, monotonically in Löwner order, for $t \to +\infty$. At any given finite time t, one has the following (ordered) generalized Schur decomposition of S_1:

$$\left(\begin{array}{c|c} Q_{11}(t)^\mathsf{T} & Q_{21}(t)^\mathsf{T} \\ \hline Q_{12}(t)^\mathsf{T} & Q_{22}(t)^\mathsf{T} \end{array}\right)\left(\begin{array}{c|c} \theta' & 0_{n\times n} \\ \hline Q(\delta(t)) & I_n \end{array}\right)\left(\begin{array}{c|c} Z_{11}(t) & Z_{12}(t) \\ \hline Z_{21}(t) & Z_{22}(t) \end{array}\right) = \left(\begin{array}{c|c} R_{11}(t) & \star \\ \hline 0_{n\times n} & R_{22}(t) \end{array}\right), \tag{7.2}$$

with abuse of notation regarding Q, *i.e.*, $Q(t)$ versus $Q(\delta(t))$. As $Q(t), Z(t) \in \mathsf{O}(2n,\mathbb{R})$, $\|Q(t)\|_\infty \leq 1$, $\|Z(t)\|_\infty \leq 1$ for all $t \in \mathbb{R}_{\geq 0}$. Then, since $Q(\delta(t))$ decreases monotonically with increasing t, there is a T such that for all $t \geq T$, $Q(T)^\mathsf{T} S_1(\delta(t))Z(T)$ is quasi upper-triangular up to machine precision. A similar argument can be made regarding $S_2(\delta(t))$. Moreover, "*deflation*" is commonly applied to the subdiagonal elements of S_1, that is, small elements are set to 0 [Kre05, Section 2.3.4]. Differently put, for sufficiently large t one either lands at the Hessenberg-triangular reduction step *cf.* [GL13, Algorithm 7.7.1], or at the final QZ step, both up to numerical precision.

Now, let us assume that $\min_i\{\lambda_i(\theta')\} \gg \mu$, such that the diagonal of $R(t)$ only contains terms sufficiently far away from 0. Under this assumption no structural changes can occur.

The QZ algorithm might proceed to remove excess terms using $2n \times 2n$-dimensional Givens rotations[12], that is, matrices of the form

$$\begin{pmatrix} I_{i-1} & 0 & 0 & 0 & 0 \\ 0 & \cos(\phi) & 0 & \sin(\phi) & 0 \\ 0 & 0 & I_{j-i-1} & 0 & 0 \\ 0 & -\sin(\phi) & 0 & \cos(\phi) & 0 \\ 0 & 0 & 0 & 0 & I_{2n-j} \end{pmatrix}, \tag{7.3}$$

for $i < j$ and $\phi \in [-\pi/2, \pi/2)$ [Kre05, p. 77], [GL13, Section 5.1.8]. However, now see that for some $\epsilon > 0$ close to 0, we have

$$\begin{pmatrix} \cos(\phi) & \sin(\phi) \\ -\sin(\phi) & \cos(\phi) \end{pmatrix}\begin{pmatrix} a & b \\ \epsilon & c \end{pmatrix} = \begin{pmatrix} \sqrt{a^2+\epsilon^2} & \star \\ 0 & \star \end{pmatrix}$$

by selecting ϕ such that $\cos(\phi) = a/(\sqrt{a^2+\epsilon^2})$, $\sin(\phi) = \epsilon/(\sqrt{a^2+\epsilon^2})$. As such, for $\epsilon \approx 0$, the rotation is close to I_2. Summarizing, not only does this procedure largely[13] preserve the structure of $(Q(t), Z(t))$ for $t \to +\infty$, we also reduce the computational cost as the QZ algorithm will be eventually initialized close to the solution, which is particularly exacerbated for larger n. This scheme is summarized as Algorithm 2.

Also, although we will not embark on this, but to deal with large data-streams we ought it imperative to remark that one can implement (1.2) recursively, *e.g.*, see [Kum90].

[12] Depending on conventions, one might consider the transpose of (7.3).

[13] We do potentially break the symplectic structure.

Algorithm 2 Numerical reverse I-projection of θ' (pseudo-Julia code).

```
Input: Iteration limit maxIter$\in \mathbb{N}$, monotonically decreasing sequence $(\delta_k)_{k\leq \texttt{maxIter}} \subset \mathbb{R}_{>0}$,
    $Q \in C^{r\geq 1}(\mathbb{R}_{>0}, \mathcal{S}^n_{\succ 0})$, $\theta' \in \mathbb{R}^{n\times n}$, $S_w \in \mathcal{S}^n_{\succ 0}$, stopping condition parameter $\epsilon > 0$.
Set Qtemp $= I_{2n}$, Ztemp $= I_{2n}$, $k = 0$.
while (stopping condition $> \epsilon$ and $k \leq$ maxIter) do
    $S_1 =$ Qtemp$^{\mathsf{T}} S_1(\delta_k, Q(\delta_k), \theta')$Ztemp, $S_2 =$ Qtemp$^{\mathsf{T}} S_2(\delta_k, \theta', S_w)$Ztemp.
    $[P_{\delta_k}, \theta_{\delta_k},$ QZ_reord$]$ = Algorithm 1 $(\delta_k, Q(\delta_k), \theta', S_w, S_1, S_2)$.
    Qtemp $\leftarrow$ QZ_reord.Q, Ztemp $\leftarrow$ QZ_reord.Z, $k \leftarrow k+1$.
end while
Output: $\theta^\star_{\delta_k}$.
```

7.2.1 Symplectic balancing As alluded to before, the balancing proposed above using does *not* necessarily preserve the underlying symplectic structure of the problem[14]. Preserving this structure can be beneficial when the spectrum $\mathrm{spec}(S_1, S_2)$ contains elements close to $\partial\mathbb{D}_1$, in that case, the symplectic structure enforces pairs of stable- and unstable eigenvalues.

Due to the block-structure of elements in $\mathsf{Sp}(2n, \mathbb{R})$ one can conveniently characterize all elements of $\mathsf{O}(2n, \mathbb{R}) \cap \mathsf{Sp}(2n, \mathbb{R}) = \mathsf{OSp}(2n, \mathbb{R})$ by

$$\mathsf{OSp}(2n, \mathbb{R}) = \left\{ \begin{pmatrix} A & B \\ -B & A \end{pmatrix} \in \mathbb{R}^{2n\times 2n} : A^{\mathsf{T}}B = B^{\mathsf{T}}A,\ A^{\mathsf{T}}A + B^{\mathsf{T}}B = I_n \right\}. \tag{7.4}$$

Reconsidering (7.2) under the standing assumption that the first n colums of n span $\mathcal{X}^s$, it seems beneficial to preserve those n columns and preferably R_{11}. Let the pair (Q, Z) correspond to the generalized Schur decomposition of the pair (S_1, S_2), then, given the structure of $\mathsf{OSp}(2n, \mathbb{R})$, the corresponding symplectic matrices $\widetilde{Q}$ and $\widetilde{Z}$ are fixed and given by

$$\widetilde{Q} = \begin{pmatrix} Q_{11} & -Q_{21} \\ Q_{21} & Q_{11} \end{pmatrix}, \quad \widetilde{Z} = \begin{pmatrix} Z_{11} & -Z_{21} \\ Z_{21} & Z_{11} \end{pmatrix}.$$

Here, we appeal to the *Laub-Mehrmann-trick*, *cf.* [Meh88]. By exploiting that $Z_{21}Z_{11}^{-1} \succ 0$, it follows that $\widetilde{Z}^{\mathsf{T}}\widetilde{Z} = I_n$ and $\widetilde{Z}^{\mathsf{T}}\Omega\widetilde{Z} = \Omega$. As such, $\tilde{Z} \in \mathsf{OSp}(2n, \mathbb{R})$ indeed. In general, given a symplectic pair (S_1, S_2), then, $(Q^{\mathsf{T}}S_1Z, Q^{\mathsf{T}}S_2Z)$ is a symplectic pair for any $Z \in \mathsf{OSp}(2n, \mathbb{R})$ and $Q \in \mathsf{O}(2n, \mathbb{R})$ since

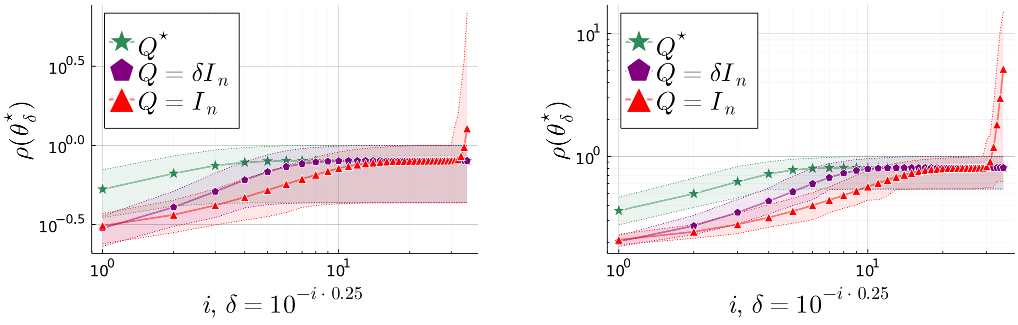
$$(Q^{\mathsf{T}}S_1Z)\Omega(Q^{\mathsf{T}}S_1Z)^{\mathsf{T}} = Q^{\mathsf{T}}S_1\Omega S_1^{\mathsf{T}}Q = Q^{\mathsf{T}}S_2\Omega S_2^{\mathsf{T}}Q = (Q^{\mathsf{T}}S_2Z)\Omega(Q^{\mathsf{T}}S_2Z)^{\mathsf{T}}.$$

With these observation in mind, one can adapt Algorithm 2, that is, by using the pair $(\widetilde{Q}, \widetilde{Z})$ instead of (Q, Z).

7.3 Numerical experiments In this section we showcase the effect of the previous analysis on the computation of $\theta^\star_\delta$. Colloquially speaking, the ideal behaviour should display: (a) fast *convergence* of $\theta^\star_{\delta_k}$ for $\delta_k \downarrow 0$, so that choice of $\{\delta_k\}_{k\geq 0} \subset \mathbb{R}_{>0}$ is not critical; (b) fast *computation* of $\theta^\star_\delta$, with the computational time invariant under the choice of $\delta > 0$; (c) convergence and computation without numerical instabilities, if $\delta_1 \approx \delta_2$, then always $\theta^\star_{\delta_1} \approx \theta^\star_{\delta_2}$. Throughout we use the shorthand notation $Q^\star = 2^{-1}\delta^2\theta'^{\mathsf{T}}(2S_w)^{-1}\theta'$ and $Q_\star = 2\delta\theta'^{\mathsf{T}}S_w\theta'$.

[14] The most simple example of a matrix in $\mathsf{O}(2n, \mathbb{R}) \setminus \mathsf{Sp}(2n, \mathbb{R})$ would be a reflection of the form

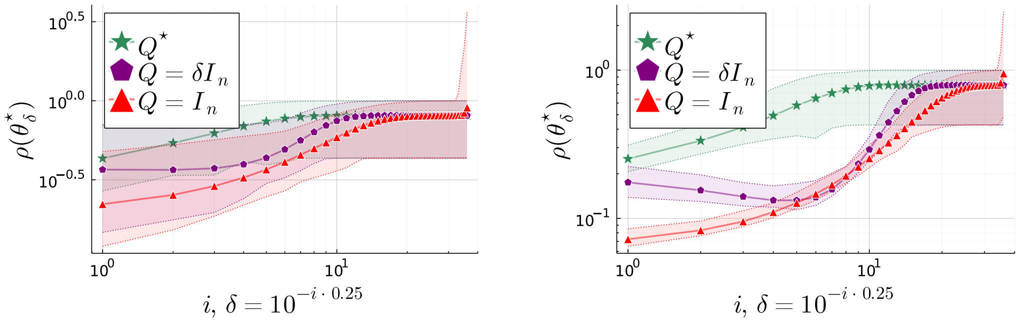
$$Z = \begin{pmatrix} -I_n & 0_{n\times n} \\ 0_{n\times n} & I_n \end{pmatrix}.$$



(a) $n = 10$

(b) $n = 100$

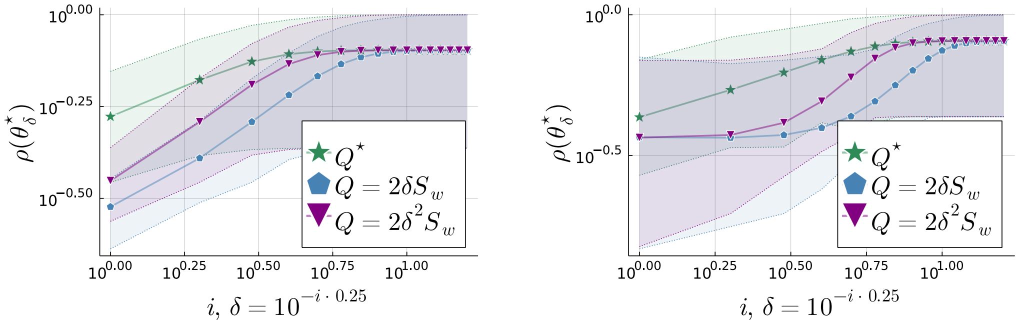
Figure 7.1: Numerical experiments (250 per δ), computing $\theta^\star_\delta$ for $\mathrm{vec}(\theta') \sim \mathcal{N}(0, I_{n^2})$ under different choices of Q and $S_w = I_n$. Each figure displays all available data.



(a) $n = 10$

(b) $n = 100$

Figure 7.2: Numerical experiments (250 per δ), computing $\theta^\star_\delta$ for $\mathrm{vec}(\theta') \sim \mathcal{N}(0, I_{n^2})$ under different choices of Q and $S_w = L^\mathsf{T} L$ for $\mathrm{vec}(L) \sim \mathcal{N}(0, I_{n^2})$. Each figure displays all available data.

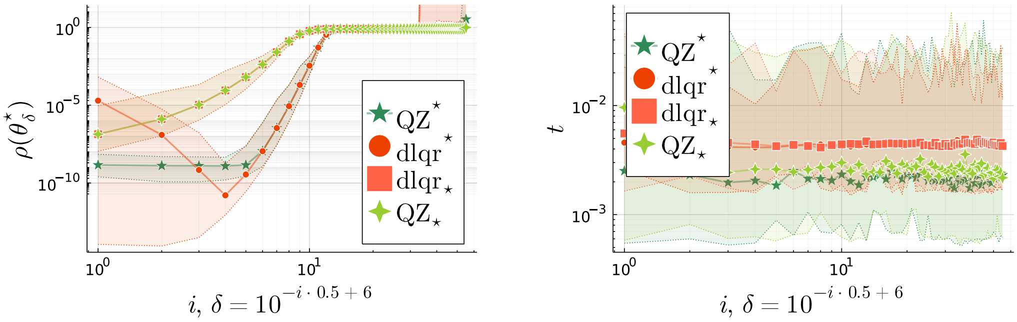


(a) $n = 10$

(b) $n = 10$

Figure 7.3: Numerical experiments (250 per δ), computing $\theta^\star_\delta$ for $\mathrm{vec}(\theta') \sim \mathcal{N}(0, I_{n^2})$ under different choices of Q and (a) $S_w = I_n$ or (b) $S_w = L^\mathsf{T} L$ for $\mathrm{vec}(L) \sim \mathcal{N}(0, I_{n^2})$. Each figure displays all available data.

(i) (Basic scaling) In the first experiment we only employ the QZ method, *i.e.*, Algorithm 1, under a range of $\delta > 0$ and different choices of $Q(\delta)$. Figure 7.1, displays the results for a fixed S_w, whereas Figure 7.2 shows the average performance under different S_w, these



(a) Convergence for $n = 10$

(b) Time to compute $\theta^\star_\delta$ for $n = 10$

Figure 7.4: Numerical experiments (250 per δ), computing $\theta^\star_\delta$ by means of the QZ algoritm or Julia's $\mathsf{dlqr}(\cdot)$ routine, for $\mathrm{vec}(\theta') \sim \mathcal{N}(0, I_{n^2})$ under different choices of Q and $S_w = (1/n^2)I_n$. Each figure displays all available data.

random S_w are by construction potentially ill-conditioned. Overall, the scaling as derived in Section 6.1 outperforms the other methods. One might argue that the aforementioned observations rely on the difference in δ, that is, $O(\delta^2)$ versus $O(\delta)$ and $O(1)$. To that end we show in Figure 7.3 how $Q^\star$ compares to for instance $Q(\delta) = 2\delta^2 S_w$. Indeed, the conditioning as derived in Section 6.1 remains beneficial.

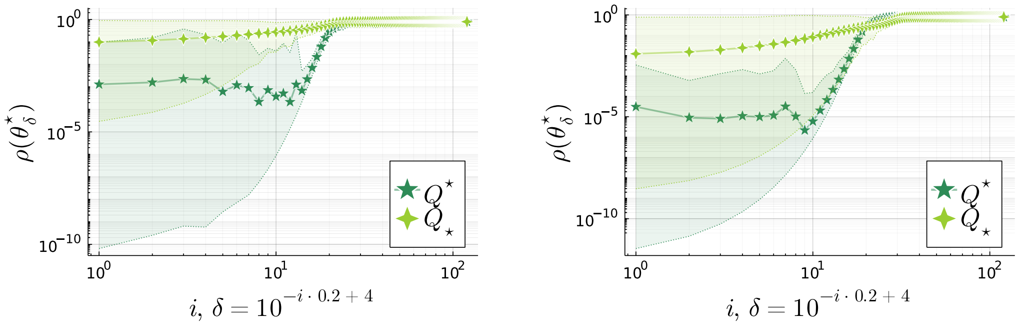
(ii) (QZ versus dlqr) The second experiment emphasizes this observation. Here, we compare the QZ method under $Q^\star$ and $Q_\star$ with the standard dlqr routine from Julia, also under $Q^\star$ and $Q_\star$, denoted $\mathsf{dlqr}^\star$ and $\mathsf{dlqr}_\star$, respectively. Figure 7.4 shows that the QZ method is well-conditioned, in contrast to the standard routine, and faster. We remark that when dlqr would fail to compute $\theta^\star_\delta$ we set $\theta^\star_\delta = +\infty I_n$. In particular, we only showcase the experiments for $n = 10$ as the for higher dimensions the dlqr routine fails too frequently.

(iii) ($Q^\star$ versus $Q_\star$) Most of the previous experiments consider relatively large S_w and prohibit the selection of $Q_\star$. In practice, however, it is unlikely that $S_w = O(I_n)$, we expect S_w to be orders of magnitude smaller. Under such an assumption we compare $Q^\star$ and $Q_\star$ in Figure 7.5-7.6. Indeed, for sufficiently small S_w, $Q_\star$ outperforms $Q^\star$. Concurrently, Figure 7.5 is shown to convey how small δ can be while having a numerically stable algorithm, in these experiments we could use $\delta = 10^{-20}$, which is far below what standard dlqr routines would allow for. Overall we see that for "*large*" S_w one could employ $Q^\star$ whereas for "*small*" S_w, $Q_\star$ is recommended.

(iv) (QZ versus iteratively balanced QZ) Here we show how the iterative method from Section 7.2 improves upon a QZ method that does not exploit prior knowledge, *i.e.*, a method that simply computes $\theta^\star_\delta$ given the 4-tuple $(\delta, Q(\delta), \theta', S_w)$. In Figure 7.7-7.8 we compare these two approaches under $Q_\star$. One can observe that the iterative method is indeed faster and the scenario as sketched in Example 7.2 is avoided, in contrast to the standard approach.

(v) (Symplectic balancing) At last, we show that for somewhat ill-conditioned problems, the symplectic balancing from Section 7.2.1 can be beneficial, see Figure 7.9.

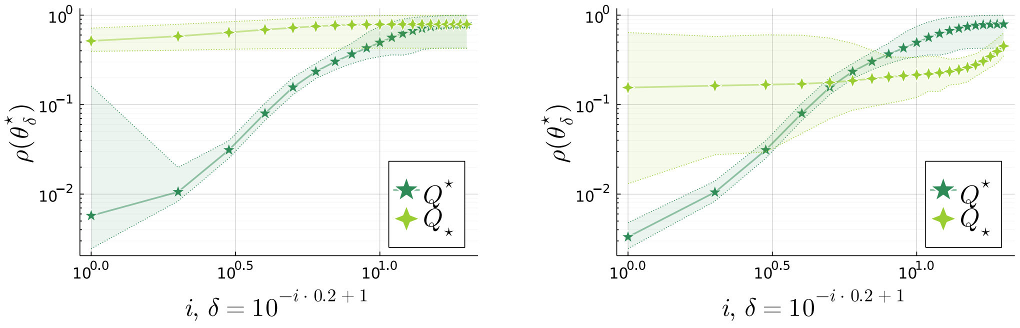
All numerical experiments were performed using Julia [Bez+17] on a i7-8550U CPU laptop with 16Gb RAM.



(a) $n = 10$

(b) $n = 10$

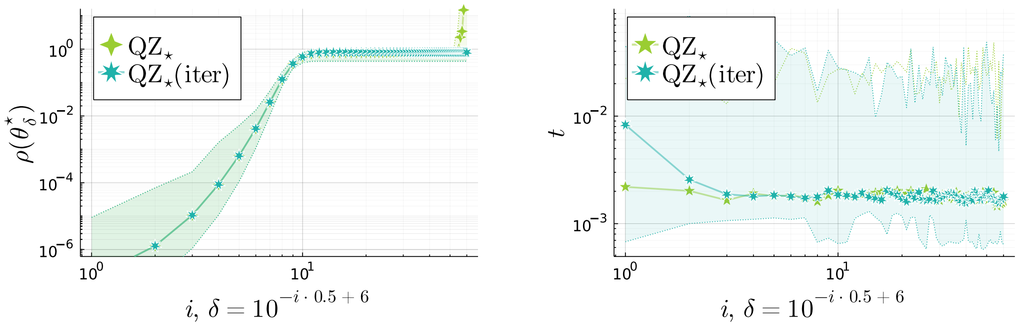
Figure 7.5: Numerical experiments (250 per δ), computing $\theta^\star_\delta$ for $\mathrm{vec}(\theta') \sim \mathcal{N}(0, I_{n^2})$ under different choices of Q and (a) $S_w = (1/n^2)L^\mathsf{T}L$ or (b) $S_w = L^\mathsf{T}L$ for $\mathrm{vec}(L) \sim \mathcal{N}(0, I_{n^2})$. Each figure displays all available data.



(a) $n = 100$

(b) $n = 100$

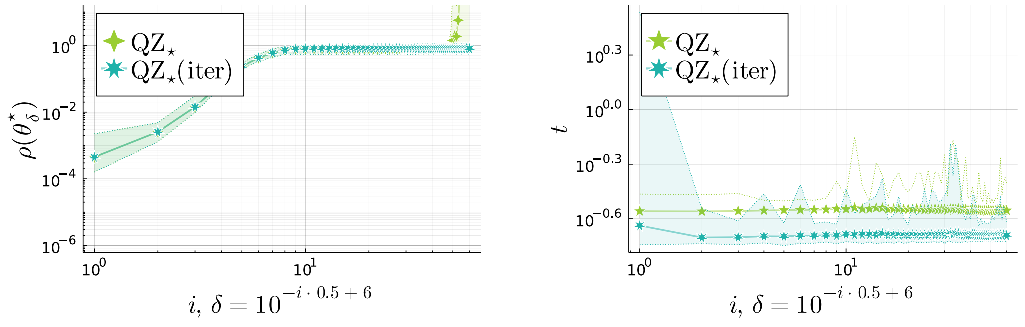
Figure 7.6: Numerical experiments (250 per δ), computing $\theta^\star_\delta$ for $\mathrm{vec}(\theta') \sim \mathcal{N}(0, I_{n^2})$ under different choices of Q and (a) $S_w = (1/n^2)L^\mathsf{T}L$ or (b) $S_w = L^\mathsf{T}L$ for $\mathrm{vec}(L) \sim \mathcal{N}(0, I_{n^2})$. Each figure displays all available data.



(a) Convergence for $n = 10$

(b) Time to compute $\theta^\star_\delta$ for $n = 10$

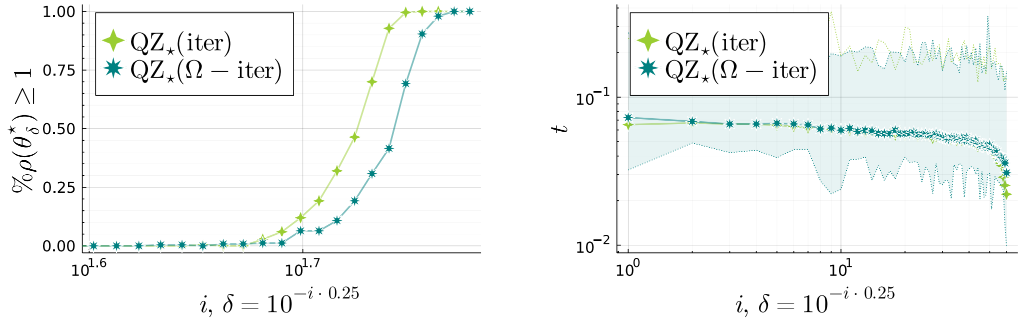
Figure 7.7: Numerical experiments (250 per δ), computing $\theta^\star_\delta$ using either the standard QZ algorithm (QZ) or the iterative scheme from Section 7.2 (QZ(iter)), for $\mathrm{vec}(\theta') \sim \mathcal{N}(0, I_{n^2})$ under $Q_\star$ and $S_w = (1/n^2)I_n$. Each figure displays all available data.



(a) Convergence for $n = 100$

(b) Time to compute $\theta^\star_\delta$ for $n = 100$

Figure 7.8: Numerical experiments (250 per δ), computing $\theta^\star_\delta$ using either the standard QZ algorithm (QZ) or the iterative scheme from Section 7.2 (QZ(iter)), for $\mathrm{vec}(\theta') \sim \mathcal{N}(0, I_{n^2})$ under $Q_\star$ and $S_w = (1/n^2)I_n$. Each figure displays all available data.



(a) Spectral errors for $n = 50$

(b) Time to compute $\theta^\star_\delta$ for $n = 50$

Figure 7.9: Numerical experiments (250 per δ), computing $\theta^\star_\delta$ by means of Algorithm 2 (QZ$_\star$(iter)) and Algorithm 2 under symplectic balancing from Section 7.2.1 (QZ$_\star$($\Omega -$ iter)). This, for $\theta' = I_n + \delta \cdot \Delta$ with $\mathrm{vec}(\Delta) \sim \mathcal{N}(0, I_{n^2})$, under $Q_\star$ and $S_w = (1/n^2)I_n$. Each figure displays all available data.

8 Conclusion and future work

This note elaborates on computational aspects of the work in [JSK23], *cf.* see Section 2. Exploiting the symplectic structure of the underlying LQR problem, a numerically appropriate algorithm, together with a selection of the cost matrix Q, is proposed. Numerical experiments show that under these cost matrices, the algorithm outperforms standard LQR routines and the sub-optimal selection of $Q = I_n$, as initially proposed in [JSK23].

Taking a continuous-time perspective akin to [KLS16] is left for future work. With respect to Section 7.1.2, another topic of interest is to formalize the improved performance due to the iterative QZ approach sketched in Section 7.2. In particular, explicit error bounds are envisioned.

Bibliography

[AL84] W. F. Arnold and A. J. Laub. "Generalized eigenproblem algorithms and software for algebraic Riccati equations". *Proceedings of the IEEE* 72.12 (1984), pp. 1746–1754.

[AM08] R. Abraham and J. E. Marsden. *Foundations of Mechanics*. American Mathematical Society, 2008.

[And+99] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. SIAM, 1999.

[Arn10] V. I. Arnold. *Mathematical Methods of Classical Mechanics*. Springer, 2010.

[BBM20] S. Biagi, A. Bonfiglioli, and M. Matone. “On the Baker-Campbell-Hausdorff Theorem: non-convergence and prolongation issues”. *Linear and Multilinear Algebra* 68.7 (2020), pp. 1310–1328.

[Ber05] D. P. Bertsekas. *Dynamic Programming and Optimal Control.* Volume 1. Athena Scientific, 2005.

[Ber07] D. P. Bertsekas. *Dynamic Programming and Optimal Control.* Volume 2. Athena Scientific, 2007.

[Bez+17] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. “Julia: A fresh approach to numerical computing”. *SIAM Review* 59.1 (2017), pp. 65–98.

[BF11] P. Benner and H. Faßbender. “On the numerical solution of large-scale sparse discrete-time Riccati equations”. *Advances in Computational Mathematics* 35 (Nov. 2011), pp. 119–147.

[BF98] P. Benner and H. Faßbender. “The symplectic eigenvalue problem, the butterfly form, the SR algorithm, and the Lanczos method”. *Linear algebra and its applications* 275 (1998), pp. 19–47.

[BGS08] B. Boots, G. J. Gordon, and S. M. Siddiqi. “A constraint generation approach to learning stable linear dynamical systems”. *Neural Information Processing Systems.* 2008, pp. 1329–1336.

[BIM11] D. A. Bini, B. Iannazzo, and B. Meini. *Numerical Solution of Algebraic Riccati Equations.* Society of Industrial and Applied Mathematics, 2011.

[BLM97] P. Benner, A. Laub, and V. Mehrmann. “Benchmarks for the numerical solution of algebraic Riccati equations”. *IEEE Control Systems* 17 (1997), pp. 18–28.

[BLW91] S. Bittanti, A. J. Laub, and J. C. Willems, eds. *The Riccati Equation.* Springer-Verlag, 1991.

[Bye87] R. Byers. “Solving the algebraic Riccati equation with the matrix sign function”. *Linear Algebra and its Applications* 85 (1987), pp. 267 –279.

[CGS20] N. Choudhary, N. Gillis, and P. Sharma. “On approximating the nearest Ω-stable matrix”. *Numerical Linear Algebra with Applications* 27.3 (2020).

[CMN12] F. Casas, A. Murua, and M. Nadinic. “Efficient computation of the Zassenhaus formula”. *Computer Physics Communications* 183.11 (2012), pp. 2386–2391.

[DK99] J. Duistermaat and J. Kolk. *Lie Groups.* Springer, 1999.

[DZ09] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Springer, 2009.

[Fat69] A. Fath. “Computational aspects of the linear optimal regulator problem”. *IEEE Transactions on Automatic Control* 14.5 (1969), pp. 547–550.

[GKL92] T. Gudmundsson, C. Kenney, and A. J. Laub. “Scaling of the discrete-time algebraic Riccati equation to enhance stability of the Schur solution method”. *IEEE Transactions on Automatic Control* 37.4 (1992), pp. 513–518.

[GKS19] N. Gillis, M. Karow, and P. Sharma. “Approximating the nearest stable discrete-time system”. *Linear Algebra and its Applications* 573 (2019), pp. 37–53.

[GL13] G. H. Golub and C. F. van Loan. *Matrix Computations.* John Hopkins University Press, 2013.

[GL86] J. D. Gardiner and A. J. Laub. “A generalization of the matrix-sign-function solution for algebraic Riccati equations”. *International Journal of Control* 44.3 (1986), pp. 823–832.

[GL91] J. D. Gardiner and A. J. Laub. “Parallel algorithms for algebraic Riccati equations”. *International Journal of Control* 54.6 (1991), pp. 1317–1333.

[Hol08] F. den Hollander. *Large Deviations.* American Mathematical Society, 2008.

[JK21] W. Jongeneel and D. Kuhn. “On Topological Equivalence in Linear Quadratic Optimal Control”. *European Control Conference.* 2021, pp. 2002–2007.

[Jos11] J. Jost. *Riemannian Geometry and Geometric Analysis.* Springer, 2011.

[JP19] Y. Jedra and A. Proutiere. “Sample complexity lower bounds for linear system identification”. *IEEE Conference on Decision and Control.* IEEE. 2019, pp. 2676–2681.

[JP20] Y. Jedra and A. Proutiere. “Finite-time identification of stable linear systems optimality of the least-squares estimator”. *IEEE Conference on Decision and Control.* IEEE. 2020, pp. 996–1001.

[JSK23] W. Jongeneel, T. Sutter, and D. Kuhn. “Efficient Learning of a Linear Dynamical System With Stability Guarantees”. *IEEE Transactions on Automatic Control* 68.5 (2023), pp. 2790–2804.

[Kat95] T. Kato. *Perturbation Theory for Linear Operators.* Springer, 1995.

[KLS16] Y.-C. Kuo, W.-W. Lin, and S.-F. Shieh. “Structure-preserving flows of symplectic matrix pairs”. *SIAM Journal on Matrix Analysis and Applications* 37.3 (2016), pp. 976–1001.

[Kre05] D. Kressner. *Numerical Methods for General and Structured Eigenvalue Problems.* Vol. 46. Lecture Notes in Computational Science and Engineering. Springer, 2005.

[Kum90] P. Kumar. "Convergence of adaptive control schemes using least-squares parameter estimates". *IEEE Transactions on Automatic Control* 35.4 (1990), pp. 416–424.

[Lau79] A. Laub. "A Schur method for solving algebraic Riccati equations". *IEEE Transactions on Automatic Control* 24.6 (1979), pp. 913–921.

[Lax07] P. D. Lax. *Linear Algebra and its Applications.* Wiley, 2007.

[LB02] S. L. Lacy and D. S. Bernstein. "Subspace identification with guaranteed stability using constrained optimization". *American Control Conference.* Vol. 4. 2002, 3307–3312 vol.4.

[LB03] S. L. Lacy and D. S. Bernstein. "Subspace identification with guaranteed stability using constrained optimization". *IEEE Transactions on Automatic Control* 48.7 (2003), pp. 1259–1263.

[Lee13] J. M. Lee. *Introduction to Smooth Manifolds.* Springer, 2013.

[LLP99] L.-Z. Lu, W.-W. Lin, and C. E. Pearce. "An efficient algorithm for the discrete-time algebraic Riccati equation". *IEEE Transactions on Automatic Control* 44.6 (1999), pp. 1216–1220.

[LR95] P. Lancaster and L. Rodman. *Algebraic Riccati Equations.* Oxford University Press, 1995.

[LX06] W.-W. Lin and S.-F. Xu. "Convergence analysis of structure-preserving doubling algorithms for Riccati-type matrix equations". *SIAM Journal on Matrix Analysis and Applications* 28.1 (2006), pp. 26–39.

[Mac95] J. Maciejowski. "Guaranteed stability with subspace methods". *Systems & Control Letters* 26.2 (1995), pp. 153 –156.

[MBO97] J. Moro, J. V. Burke, and M. L. Overton. "On the Lidskii–Vishik–Lyusternik perturbation theory for eigenvalues of matrices with arbitrary Jordan structure". *SIAM Journal on Matrix Analysis and Applications* 18.4 (1997), pp. 793–817.

[Meh+09] C. Mehl, V. Mehrmann, A. C. Ran, and L. Rodman. "Perturbation analysis of Lagrangian invariant subspaces of symplectic matrices". *Linear and Multilinear Algebra* 57.2 (2009), pp. 141–184.

[Meh88] V. Mehrmann. "A symplectic orthogonal method for single input or single output discrete time optimal quadratic control problems". *SIAM Journal on Matrix Analysis and Applications* 9.2 (1988), pp. 221–247.

[MS73] C. B. Moler and G. W. Stewart. "An algorithm for generalized matrix eigenvalue problems". *SIAM Journal on Numerical Analysis* 10.2 (1973), pp. 241–256.

[NP20] Y. Nesterov and V. Y. Protasov. "Computing closest stable nonnegative matrix". *SIAM Journal on Matrix Analysis and Applications* 41.1 (2020), pp. 1–28.

[NS90] H. Nijmeijer and A. van der Schaft. *Nonlinear Dynamical Control Systems.* Springer, 1990.

[ONV13] F.-X. Orbandexivry, Y. Nesterov, and P. Van Dooren. "Nearest stable system using successive convex approximations". *Automatica* 49.5 (2013), pp. 1195 –1203.

[PLS80] T. Pappas, A. Laub, and N. Sandell. "On the numerical solution of the discrete-time algebraic Riccati equation". *IEEE Transactions on Automatic Control* 25.4 (1980), pp. 631–641.

[Sim+18] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht. "Learning Without Mixing: Towards A Sharp Analysis of Linear System Identification". *Conference On Learning Theory.* 2018, pp. 439–473.

[SM16] F. Sosa and J. Moro. "First order asymptotic expansions for eigenvalues of multiplicatively perturbed matrices". *SIAM Journal on Matrix Analysis and Applications* 37.4 (2016), pp. 1478–1504.

[SMM20] F. Sosa, J. Moro, and C. Mehl. "First order structure-preserving perturbation theory for eigenvalues of symplectic matrices". *SIAM Journal on Matrix Analysis and Applications* 41.2 (2020), pp. 657–690.

[SR19] T. Sarkar and A. Rakhlin. "Near optimal finite time identification of arbitrary linear dynamical systems". *International Conference on Machine Learning.* 2019, pp. 5610–5618.

[SRD20] T. Sarkar, A. Rakhlin, and M. A. Dahleh. "Nonparametric Finite Time LTI System Identification". *arXiv e-prints arXiv:1902.01848* (2020).

[Sun98] J.-G. Sun. "Perturbation theory for algebraic Riccati equations". *SIAM Journal on Matrix Analysis and Applications* 19.1 (1998), pp. 39–65.

[Tur+13] K. Turksoy, E. S. Bayrak, L. Quinn, E. Littlejohn, and A. Cinar. "Guaranteed stability of recursive multi-input-single-output time series models". *American Control Conference.* 2013, pp. 77–82.

[Ume+18] J. Umenberger, J. Wågberg, I. R. Manchester, and T. B. Schön. "Maximum likelihood identification of stable linear dynamical systems". *Automatica* 96 (2018), pp. 280 –292.

[Van+00] T. Van Gestel, J. A. K. Suykens, P. Van Dooren, and B. De Moor. "Imposing stability in subspace identification by regularization". *IEEE Conference on Decision and Control.* 2000, 1555–1560 vol.2.

[Van+01] T. Van Gestel, J. A. K. Suykens, P. Van Dooren, and B. De Moor. "Identification of stable models in subspace identification by using regularization". *IEEE Transactions on Automatic Control* 46.9 (2001), pp. 1416–1420.

[Var84] V. S. Varadarajan. *Lie groups, Lie algebras, and their representations*. Springer, 1984.

[VD81] P. Van Dooren. "A generalized eigenvalue approach for solving Riccati equations". *SIAM Journal on Scientific and Statistical Computing* 2.2 (1981), pp. 121–135.

[VODM96] P. Van Overschee and B. De Moor. *Subspace identification for linear systems: Theory-Implementation-Applications*. Kluwer Academic Publishers, 1996.

[War81] R. C. Ward. "Balancing the generalized eigenvalue problem". *SIAM Journal on Scientific and Statistical Computing* 2.2 (1981), pp. 141–152.

[Wil71] J. Willems. "Least squares stationary optimal control and the algebraic Riccati equation". *IEEE Transactions on Automatic Control* 16.6 (1971), pp. 621–634.