Stable Linear System Identification With Prior Knowledge by Riemannian Sequential Quadratic Optimization

Mitsuaki Obara, Kazuhiro Sato, Member, IEEE, Hiroki Sakamoto, Takayuki Okuno, and Akiko Takeda

Abstract—We consider an identification method for a linear continuous time-invariant autonomous system from noisy state observations. In particular, we focus on the identification to satisfy the asymptotic stability of the system with some prior knowledge. To this end, we propose to model this identification problem as a Riemannian nonlinear optimization (RNLO) problem, where the stability is ensured through a certain Riemannian manifold and the prior knowledge is expressed as nonlinear constraints defined on this manifold. To solve this RNLO, we apply the Riemannian sequential quadratic optimization (RSQO) that was proposed by Obara, Okuno, and Takeda (2022) most recently. RSQO performs quite well with theoretical guarantee to find a point satisfying the Karush–Kuhn–Tucker conditions of RNLO. In this article, we demonstrate that the identification problem can be indeed solved by RSQO more effectively than competing algorithms.

Index Terms—Riemannian nonlinear optimization (RNLO), Riemannian sequential quadratic optimization (RSQO), stable linear system, system identification.

I. INTRODUCTION

System identification of a linear continuous time-invariant autonomous system

\[ \dot{x}(t) = Ax(t) \]  

(1)

with the state vector \( x(t) \in \mathbb{R}^n \) is a task to estimate \( A \in \mathbb{R}^{n \times n} \) from measured state data and is one of the most important topics to predict the future state of a real system. System identification has been investigated for many years in several settings. For example, prediction error methods [1], [2], [3], [4], [5] and subspace identification methods [6], [7], [8], [9], [10] are conventional identification methods for linear time-invariant state space systems. Dynamic mode decomposition (DMD) deals with the system of the particular form (1), which can be possibly large-scale one such as fluid flows [11], [12].

In real applications, it is often crucial to identify a system satisfying the asymptotic stability. Yet, it is completely nontrivial to ensure the stability in the identification methods such as those above; e.g., DMD spectrum has been shown to be considerably sensitive to measurement noise and hence, its resultant system may be unstable [13]. Several approaches have been investigated to overcome the difficulty: subspace identification methods with guaranteed stability have been proposed in [14], [15], and [16]. In [17], a constraint generation approach has been proposed. A Lagrange relaxation is used to ensure several types of stability including the asymptotic stability for linear time-invariant state space models [18] as well as a different stability called the global incremental \( \ell^2 \) stability for nonlinear systems [19], [20]. It is worth mentioning that, among the above identification methods, those presented in [14], [15], [16], [17], [18], [19], and [20] are based on convex optimization.

In addition to the stability, another important characteristic in the identification is prior knowledge. As described in [1], Chapter 16, identification should reflect the prior knowledge peculiar to the system, such as the nonnegativity of all or partial components of the system \( A \). Nevertheless, the conventional identification methods are not able to generate a stable matrix \( A \) with prior knowledge information. Indeed, to impose the knowledge, it is necessary to deal with their nonconvexity, which requires additional techniques from nonconvex optimization.

In this article, we propose a new method for identifying a system, from noisy data, that satisfies both asymptotic stability and prior knowledge. This method first discretizes the system, and then applies the prediction error method, which is formulated as solving a least squares problem. However, to identify the desired system, we must incorporate both stability and prior knowledge into this optimization problem. Motivated by the fact that an asymptotically stable linear system can be expressed in a port-Hamiltonian form [21, Proposition 1], we handle the stability of the system through a Riemannian manifold as in [22]. Meanwhile, we deal with the prior knowledge information as nonlinear constraints defined on the manifold. After all, our identification method reduces to solving a constrained nonlinear optimization problem on a Riemannian manifold, called RNLO for short.

Riemannian nonlinear optimization problem (RNLO) is one of the general optimization classes proposed in [23] and researches for algorithms to solve RNLO are in progress, e.g., [24], [25]. Remarkably, the modeling enables us to identify a stable \( A \) even when we use noisy observations, which usually spoil the stability in system identification. The modeling also allows us to take various prior knowledge into account by virtue of the generality of RNLO. To solve RNLO, we propose to apply the Riemannian sequential quadratic optimization method (RSQO) that was presented recently in [25].
is the global convergence property to Karush–Kuhn–Tucker (KKT) points of RNLO and locally fast convergence speed.

Our contributions are summarized as follows.
1) We propose a prediction error method to identify a linear system satisfying the asymptotic stability with prior knowledge from noisy state observations. The key ingredients of this method are formulating the identification problem as RNLO and solving it with RSQO.
2) We conduct numerical experiments with comparisons to demonstrate that RNLO modeling and RSQO are very promising.

Organization of Article: The rest of this article is organized as follows. In the following subsection, we introduce notations and terminologies. In Section III, we formulate the system identification problem as RNLO. In Section IV, we introduce RSQO of the form tailored to the obtained RNLO and moreover show its theoretical properties. We also exploit the geometry of the problem. In Section V, we demonstrate that the proposed formulation and RSQO are promising through experimental comparisons. Finally, Section VI concludes this article with some remarks and discussions about future work.

Notations and Terminologies: The sets of real numbers and complex ones are denoted by \( \mathbb{R} \) and \( \mathbb{C} \), respectively. Let \( \mathbb{N} \) be the set of all the natural numbers, i.e., \( \mathbb{N} := \{1, 2, 3, \ldots \} \). We denote the identity matrix of size \( n \times n \) by \( I \in \mathbb{R}^{n \times n} \). Given a matrix \( Y \in \mathbb{R}^{n \times n} \), \( Y^\top \) and \( \text{tr}(Y) \) represent the transpose of \( Y \) and the trace of \( Y \), respectively. Let \( \| \cdot \|_v \) denote the Frobenius norm of a matrix, i.e., \( \| Y \|_v := \sqrt{\text{tr}(Y^\top Y)} \) for any \( Y \in \mathbb{R}^{n \times n} \). Let \( \| \cdot \| \) denote the Euclidean norm of a vector, i.e., \( \| v \| := \sqrt{v_1^2 + \cdots + v_n^2} \) for any \( v \in \mathbb{R}^n \). Define the \( n \times n \)th standard basis of \( \mathbb{R}^n \) by \( e_i := (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^n \), where the \( i \)-th element is one and the others are zeros. Given a sufficiently smooth function \( w : \mathbb{R}^{n \times n} \rightarrow \mathbb{R} \), the Euclidean gradient of \( w \) at \( Y \in \mathbb{R}^{n \times n} \) is denoted by \( \nabla_w(Y) \). Let \( \mathcal{V} \) be any vector space and \( \mathcal{W} \in \mathcal{V} \). Denote the tangent space at \( \eta \in \mathbb{R}^{n \times n} \) by \( \mathcal{W}_\eta := \{ \eta \mathcal{W} \} \subseteq \mathbb{R}^{n \times n} \). Clearly, \( \mathcal{W} \) is a vector space and \( \mathcal{W}_\eta \) is a subspace of \( \mathbb{R}^{n \times n} \). By \( \mathcal{W} \), we denote the vector space of \( n \times n \) matrices at \( \eta \). By \( \mathcal{W} \), we denote the tangent space at \( \eta \) in \( \mathbb{R}^{n \times n} \). The tangent space at \( P \in \mathbb{R}^{n \times n} \) is canonically identified with the set of symmetric matrices denoted by \( \text{Sym}(n) \). Under \( T_{P} \text{Sym}_{++}(n) \simeq \text{Sym}(n) \), we equip \( \text{Sym}_{++}(n) \) with a Riemannian metric at \( P \in \text{Sym}_{++}(n) \) defined by

\[
\langle \xi_P, \eta_P \rangle_p := \text{tr}(P^{-1} \xi_P P^{-1} \eta_P)
\]  

(3)

for any \( \xi_P, \eta_P \in \text{Sym}(n) \). Let us define an operator \( \text{sym} : \mathbb{R}^{n \times n} \rightarrow \text{Sym}(n) \) by \( \text{sym}(Y) := \frac{Y + Y^\top}{2} \). Let \( m : \text{Sym}_{++}(n) \rightarrow \mathbb{R} \) be a twice continuously differentiable function and \( \nabla m \) be the smooth extension of \( m \) to the Euclidean space \( \mathbb{R}^{n \times n} \). The Riemannian gradient of \( m \) at \( P \) is

\[
\nabla m(P) = P \text{sym}(\nabla m(P)) P^\top
\]  

(4)

where \( \nabla m(P) \) is the Euclidean gradient of \( m \) at \( P \). On \( \text{Sym}_{++}(n) \) equipped with (3), we introduce retraction [26] by Retr\(_{P}(\xi_P) = P + \xi_P + \frac{1}{2!} \xi_P P^{-1} \xi_P \) for all \( \xi_P \in \text{Sym}(n) \). See [26] and [27] for the details of the concepts and the notations on optimization on Riemannian manifolds and the geometries of the \( \text{Skew}(n) \) and \( \text{Sym}_{++}(n) \).

B. Characterizations of Stability

Let us start with the definitions of the stability: \( A \in \mathbb{R}^{n \times n} \) is stable if the real parts of all the eigenvalues of the matrix \( A \) are negative. We say that the system (1) is asymptotically stable if \( A \) is stable.

In the following proposition, a useful result for the stability is provided. The proof can be found in the literature, e.g., [22], Section III] and [21].

Proposition 2.1: A matrix \( A \in \mathbb{R}^{n \times n} \) is stable, if and only if there exists \( (J, R, Q) \in \text{Skew}(n) \times \text{Sym}_{++}(n) \times \text{Sym}_{++}(n) \) such that

\[
A = (J - R)Q.
\]  

(5)

This proposition motivates us to consider optimization problems with respect to \( (J, R, Q) \) to ensure the stability of the system. Note that, for any stable matrix \( A \), the triplet \( (J, R, Q) \) satisfying (5) is not unique.

Define \( \text{Def} := \text{Skew}(n) \times \text{Sym}_{++}(n) \times \text{Sym}_{++}(n) \) and write \( \theta := (J, R, Q) \in \text{Def} \). Clearly, \( \text{Def} \) is a product Riemannian manifold and its tangent space at \( \theta \) is expressed as \( T_{\theta} \text{Def} \simeq \text{Skew}(n) \times \text{Sym}_{++}(n) \times \text{Sym}_{++}(n) \), on which a Riemannian metric at \( \theta \) is defined as

\[
\langle \xi_\theta, \eta_\theta \rangle_{\theta} := \text{tr}(\xi_\theta^\top \eta_\theta^\top) + \text{tr}(R^{-1} \xi_\theta R^{-1} \eta_\theta) + \text{tr}(Q^{-1} \xi_\theta Q^{-1} \eta_\theta)
\]  

for any \( \xi_\theta, \eta_\theta \in \text{Def}(\theta) \). We use the retraction \( \text{Retr}_\theta(\xi_\theta) = (J + \xi_\theta, R + \xi_\theta R^{-1} \xi_\theta, Q + \xi_\theta + \frac{1}{2} \xi_\theta Q^{-1} \xi_\theta Q^{-1} \) for all \( \xi_\theta \in \text{Def}(\theta) \). We refer readers to [27] for the geometry of a product manifold.

III. PROBLEM SETUP

In this section, we formulate optimization problems for the identification of a stable linear system with prior knowledge and derive optimality conditions.

A. Problem Formulation

In this article, we assume that a system to be identified is of the form (1) with stable \( A \). Thus, from Proposition 2.1, a discretized model with noise \( \varepsilon_k \) of the system by the Euler method is expressed by

\[
x_{k+1} = (I + hJ - RQ)x_k + \varepsilon_k
\]  

(6)

where \( h > 0 \) is the sampling interval. That is, the noise \( \varepsilon_k \) can be interpreted as the prediction error at time \( k \). From (6), we can define the least-square function

\[
f(\theta) := \frac{1}{N} \sum_{k=0}^{N-1} \| \varepsilon_k \|^2
\]
\[ = \frac{1}{N} \| X' - (I + h(J - R)Q)X \|^2, \] (7)

where \( X := (x_0, x_1, \ldots, x_{N-1}) \in \mathbb{R}^{n \times N} \) and \( X' := (x_1, x_2, \ldots, x_N) \in \mathbb{R}^{n \times N} \). Note that the system (6) is a nonlinear regression model, because there exists a multiplication of the parameters \( J - R \) and \( Q \).

Using the least-square function (7), we consider the following optimization problem:

\[
\begin{align*}
\text{minimize} & \quad f(\theta) \\
\text{subject to} & \quad g_{ij}(\theta) \leq 0, \quad ((i, j) \in \mathcal{I}), \\
& \quad h_{ij}(\theta) = 0, \quad ((i, j) \in \mathcal{E}).
\end{align*}
\] (8)

Here, \( \mathcal{I}, \mathcal{E} \subseteq \{1, \ldots, n\} \times \{1, \ldots, n\} \) are the index sets of inequality and equality constraints, respectively. Moreover, \( \{g_{ij}\}_{(i,j) \in \mathcal{I}}, \{h_{ij}\}_{(i,j) \in \mathcal{E}} : \mathcal{M} \rightarrow \mathbb{R} \) are continuously differentiable functions. The problem (8) is a Riemannian nonlinear optimization problem (RNLO) [23], i.e., a constrained nonlinear optimization problem over a Riemannian manifold. It should be noted that this formulation is novel for a stable linear system identification with prior knowledge.

Any inequality and equality constraints are acceptable in (8) as long as they are continuously differentiable with respect to \( \theta \); for example, we can deal with nonnegativity (or nonpositivity) of the elements \( c_i(J - R)Qc_j \geq 0 \) and an element equality, \( e_i(J - R)Qc_j = c_{ij} \) for a given constant \( c_{ij} \in \mathbb{R} \). In Section V, we consider box-type inequalities.

**B. Optimality Conditions**

In what follows, we define the KKT conditions and relevant concepts for RNLO. We say that \( \theta^* \in \mathcal{M} \) satisfies the KKT conditions of RNLO (8) if there exist Lagrange multipliers \( \{\mu_{ij}^*\}_{(i,j) \in \mathcal{I}} \subseteq \mathbb{R} \) and \( \{\lambda_{ij}^*\}_{(i,j) \in \mathcal{E}} \subseteq \mathbb{R} \) such that the following hold:

\[
\begin{align*}
\mathrm{grad} f(\theta^*) + \sum_{(i,j) \in \mathcal{I}} \mu_{ij}^* \mathrm{grad} g_{ij}(\theta) + \sum_{(i,j) \in \mathcal{E}} \lambda_{ij}^* \mathrm{grad} h_{ij}(\theta) &= 0, \\
\mu_{ij}^* &
\geq 0, \quad g_{ij}(\theta^*) \leq 0, \quad \mu_{ij}^* g_{ij}(\theta^*) = 0, \quad ((i, j) \in \mathcal{I}), \\
h_{ij}(\theta^*) &= 0, \quad ((i, j) \in \mathcal{E}),
\end{align*}
\] (9)

where the operator grad denotes the Riemannian gradient as in Section I. We call \( \theta^* \) a KKT point of (8). For brevity, we often write \( \mu^* \in \mathbb{R}^{\mathcal{I}} \) and \( \lambda^* \in \mathbb{R}^{\mathcal{E}} \) for \( \{\mu_{ij}^*\}_{(i,j) \in \mathcal{I}} \) and \( \{\lambda_{ij}^*\}_{(i,j) \in \mathcal{E}} \), respectively. It is known that, under some conditions called constraint qualifications, the KKT conditions are necessary ones for the optimality; i.e., a local minimizer satisfies the KKT conditions under constraint qualifications [23], [28].

**IV. OPTIMIZATION METHODS**

In this section, we introduce a specific method for solving (8), RSQO that was originally presented in [25].

**A. Geometry of Problems**

In the following theorem, we derive the Riemannian gradient of \( f \) specifically.

**Theorem 4.1:** Given \( \theta = (J, R, Q) \in \mathcal{M} \), the Riemannian gradient of \( f \) in (8) is

\[
\frac{\mathrm{grad} f(\theta)}{= (\text{skew}(\nabla_J \mathcal{T}(\theta)), R \text{sym}(\nabla_R \mathcal{T}(\theta))R, Q \text{sym}(\nabla_Q \mathcal{T}(\theta)))},
\] (10)

where \( \mathcal{T} \) is the smooth extension of \( f \) to the Euclidean space and

\[
\begin{align*}
\nabla_J \mathcal{T}(\theta) &= -\frac{2h}{N} (X' - (I + h(J - R)Q)X)X^T Q^T, \\
\nabla_R \mathcal{T}(\theta) &= -\nabla_J \mathcal{T}(\theta), \\
\nabla_Q \mathcal{T}(\theta) &= -\frac{2h}{N} (J^T - R^T)(X' - (I + h(J - R)Q)X)X^T
\end{align*}
\] (11)

are the Euclidean gradients on \( J, R, \) and \( Q \), respectively.

**Proof:** Note that the Riemannian gradient of the function on the product manifold is organized component-wisely [27]. Using this fact, we derive the Riemannian gradient of our problem. Hereafter, we abbreviate \( \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times n} \) as \( \mathbb{R}^{n \times n \times 3} \) for brevity.

Let \( \mathcal{T} \) be the smooth extension of \( f \) to the Euclidean space. We define \( \pi_k : \mathbb{R}^{n \times n \times 3} \rightarrow \mathbb{R}^n \) by

\[
\pi_k(\theta) := x_{k+1} - Zx_k,
\] (14)

where \( \theta := (J, R, Q) \in \mathbb{R}^{n \times n \times 3} \) and \( Z := I + h(J - R)Q \). By substituting (14) into (7), we have \( \mathcal{T}(\theta) = \frac{1}{N} \sum_{k=1}^{N} \pi_k(\theta) \mathcal{P}(\theta) \) and its directional derivative at \( \theta \in \mathbb{R}^{n \times n \times 3} \) along \( \mathcal{V}_\theta = (\nabla_J \mathcal{T}(\theta), \nabla_R \mathcal{T}(\theta), \nabla_Q \mathcal{T}(\theta)) \in \mathbb{R}^{n \times n \times 3} \)

\[
\begin{align*}
\mathcal{D} \mathcal{T}(\theta)\mathcal{V}_\theta &= \frac{2}{N} (\mathcal{D} \pi_k(\theta)\mathcal{V}_{\pi_k(\theta)})^T \pi_k(\theta).
\end{align*}
\] (15)

Now, we derive the explicit form of \( \mathcal{D} \pi_k(\theta)|_{\mathcal{V}_{\pi_k(\theta)}} \) in (15). To this end, we first calculate the directional derivative along \( \mathcal{V}_\theta = (\nabla_J \mathcal{T}(\theta), \nabla_R \mathcal{T}(\theta), \nabla_Q \mathcal{T}(\theta)) \), as \( \mathcal{D} \pi_k(\theta)|_{(\nabla_J \mathcal{T}(\theta), 0, 0)} = -h\nabla_J \pi_k(\theta) \). Similarly, we have \( \mathcal{D} \pi_k(\theta)|_{(0, \nabla_R \mathcal{T}(\theta), 0)} = h\nabla_R \pi_k(\theta) \) and \( \mathcal{D} \pi_k(\theta)|_{(0, 0, \nabla_Q \mathcal{T}(\theta))} = -h(J - R) \nabla_Q \pi_k(\theta) \).

Thus, the explicit form of \( \mathcal{D} \pi_k(\theta)|_{\mathcal{V}_{\pi_k(\theta)}} \) is

\[
\begin{align*}
\mathcal{D} \pi_k(\theta)|_{\mathcal{V}_{\pi_k(\theta)}} &= -h((\nabla_J \mathcal{T}(\theta), \nabla_R \mathcal{T}(\theta), \nabla_Q \mathcal{T}(\theta))x_k).
\end{align*}
\] (16)

Using the result, we derive the Riemannian gradient of \( f \). By substituting (14) and (16) into (15), we have

\[
\begin{align*}
\mathcal{D} \mathcal{T}(\theta)|_{\mathcal{V}_\theta} &= \mathrm{tr} \left( \nabla_J \mathcal{T}(\theta) \sum_{k=1}^{N} \frac{2h}{N} (x_{k+1} - Zx_k)x_k^T \mathcal{V}_\theta \right) \\
&\quad + \mathrm{tr} \left( \nabla_R \mathcal{T}(\theta) \sum_{k=1}^{N} \frac{2h}{N} (x_{k+1} - Zx_k)x_k^T \mathcal{V}_\theta \right) \\
&\quad + \mathrm{tr} \left( \nabla_Q \mathcal{T}(\theta) \sum_{k=1}^{N} \frac{2h}{N} (J^T - R^T)(x_{k+1} - Zx_k)x_k^T \mathcal{V}_\theta \right).
\end{align*}
\] (17)

Here, it follows from \( x_{k+1} = X'e_k \) and \( x_k = Xe_k \) that

\[
\sum_{k=1}^{N} (x_{k+1} - Zx_k)x_k^T = (X' - ZX) \left( \sum_{k=1}^{N} e_k e_k^\top \right) X^T = (X' - ZX) X^T,
\] (18)
where the second equality holds from \( \sum_{k=1}^{N-1} e_k e_k^\top = I \). Thus, since \( M \subseteq \mathbb{R}^{n \times n} \) holds, combining (17) and (18) provides (11), (12), and (13) for any \( \theta \in M \). By projecting (11), (12), and (13) onto \( T_\theta M \) according to (2) and (4), we obtain the Riemannian gradient of the form (10).

In a similar manner, we can derive the Riemannian gradient of constraints although we focus on that of the objective function here.

### B. Riemannian Sequential Quadratic Optimization

In this subsection, we briefly explain the RSQO for solving (8) using notations and terminologies we have set up so far. See [25, Section 3] for more details.

1) **Description of RSQO**: RSQO is an iterative method. Let \( \theta^k \in M \) be a current iterate. RSQO first solves the following quadratic program

\[
\begin{align*}
\text{minimize} & \quad \tilde{f}^k (\Delta \theta^k) \\
\text{subject to} & \quad \tilde{g}^k (\Delta \theta^k) \leq 0, \quad ((i, j) \in I), \\
& \quad \tilde{h}^k (\Delta \theta^k) = 0, \quad ((i, j) \in \mathcal{E}),
\end{align*}
\]

(19)

where

\[
\tilde{f}^k (\Delta \theta^k) := \frac{1}{2} (B^k \Delta \theta^k, \Delta \theta^k)_{\phi^k} + \langle \text{grad} f(\theta^k), \Delta \theta^k \rangle_{\phi^k},
\]

(20)

\[
\tilde{g}^k (\Delta \theta^k) := g_{ij}(\theta^k) + \langle \text{grad} g_{ij}(\theta^k), \Delta \theta^k \rangle_{\phi^k},
\]

(21)

\[
\tilde{h}^k (\Delta \theta^k) := h_{ij}(\theta^k) + \langle \text{grad} h_{ij}(\theta^k), \Delta \theta^k \rangle_{\phi^k},
\]

(22)

and \( B^k : T_{\theta^k} M \to T_{\theta^k} M \) is a symmetric positive-definite linear operator. Since the problem can be expressed as a Euclidean convex quadratic optimization problem, it can be solved by, for example, interior-point methods or active set methods. We denote the Lagrange multipliers at the solution \( \Delta \theta^k \) corresponding to the inequality and equality constraints by \( \mu^k \in \mathbb{R}^{m,M} \) and \( \lambda^k \in \mathbb{R}^{\mathcal{E},M} \), respectively.

Next, RSQO determines the step length by using the \( \ell_1 \) penalty function defined as

\[
P_{\mu^k}(\theta) := f(\theta) + \rho \left( \sum_{(i, j) \in \mathcal{I}} \max \{0, g_{ij}(\theta)\} + \sum_{(i, j) \in \mathcal{E}} |h_{ij}(\theta)| \right),
\]

where \( \rho > 0 \) is a penalty parameter [24]. RSQO first sets

\[
\rho^k = \rho^{k-1}, \quad \text{if } \rho^{k-1} \geq v^k,
\]

otherwise,

(23)

where \( v^k := \max \{ \max_{(i, j) \in \mathcal{I}} \mu^k, \max_{(i, j) \in \mathcal{E}} |\lambda^k| \} \) and \( \epsilon > 0 \) is a prescribed algorithmic parameter. Then, by using \( P_{\mu^k} \) as a merit function, we find the smallest nonnegative integer \( t \) satisfying

\[
\frac{\gamma \beta^t}{(B^t \Delta \theta^k, \Delta \theta^k)_{\phi^k}} \leq P_{\mu^k}(\theta^k) - P_{\mu^k}(\text{Retr}_{\phi^k}(\beta^t \Delta \theta^k))
\]

(24)

and set \( \alpha^k = \beta^t \). Using the search direction and the step length, RSQO updates the iterate. The above procedure is formalized as in Algorithm 1.

2) **Convergence Properties**: First, the line search terminates within finitely many trials in RSQO, because the search direction is a descent direction for \( P_{\mu^k} \circ \text{Retr}_{\phi^k} \). For details, see [25, Proposition 3.9, Remark 3.10]. Next, we show that RSQO has the global convergence property under the following assumptions, standard in nonlinear optimization:

**Assumption 1**: There exist \( m, M > 0 \) such that, for any \( k \)

\[
m \| \Delta \theta^k \|_{\phi^k}^2 \leq \langle B^k | \Delta \theta^k, \Delta \theta^k \rangle_{\phi^k} \leq M \| \Delta \theta^k \|_{\phi^k}^2
\]

for all \( \Delta \theta^k \in T_{\theta^k} M \).

**Assumption 2**: The subproblem (19) is feasible at each iteration.

**Assumption 3**: The generated sequence \( \{ (\theta^k, \mu^k, \lambda^k) \} \) is bounded. Here, the boundedness of \( \{ \theta^k \} \) is with respect to the distance induced from the Riemannian metric, while those of \( \{ \mu^k \} \) and \( \{ \lambda^k \} \) are in the sense of the Euclidean distance.

Assumptions 2 and 3 are not easy to check in prior. But, we observed that they were not violated in the numerical experiments. By Assumption 3, any accumulation point of \( \{ \theta^k \} \) stays at \( M \).

**Theorem 4.2**: [25, Theorem 3.12] Suppose Assumptions 1–3. Let \((\theta^\ast, \mu^\ast, \lambda^\ast)\) be any accumulation point of \( \{ (\theta^k, \mu^k, \lambda^k) \} \) generated by RSQO. Then, \((\theta^\ast, \mu^\ast, \lambda^\ast)\) satisfies the KKT conditions of RNLO (8).

### V. NUMERICAL SIMULATIONS

In this section, we demonstrate the effectiveness of our RNLO modeling and RSQO. We conduct numerical experiments on a random system and make comparisons. In Section V-A, we introduce a problem setting as well as other modelings for the comparisons. In Section V-B, we describe the solver settings and an evaluation index for the experiments. In Section V-C, we show the numerical results and discuss the effect of choice of the modelings and the algorithms. All the experiments are implemented in Matlab-R2023a and Manopt [29] on a Windows 10 Pro with 2.60 GHz Core i9-11980HK CPU and 64.0 GB memory.

### A. Problem Setting: Synthetic System

In this experiment, we consider the following synthetic system:

\[
\begin{align*}
\text{minimize} & \quad f(\theta) = \frac{1}{N} \| X' - (I + h(J - R)Q)X \|_F^2 \\
\text{subject to} & \quad \| i \| \leq c_i^\top (J - R)Qe_i - c_i, \\
& \quad (i, j) \in \mathcal{I} \cup \mathcal{I}_z, \\
& \quad k_j^2 \leq (e_i^\top (J - R)Qe_i - c_i)^2, \\
& \quad (i, j) \in \mathcal{I}_z,
\end{align*}
\]

(25a)

(25b)

(25c)
where \( I_1, I_2 \) are disjoint subsets of the whole indices \( \{1, \ldots, n\} \times \{1, \ldots, n\} \). \( I_1 \) is the index set of one-box constraints, for each \((i, j)\) of which the \(i\)th component of the true system belongs to \([l_{ij}, r_{ij}] \subseteq \mathbb{R}\). Similarly, \( I_2 \) is the one of two-box constraints, for each \((i, j)\) of which the true \((i, j)\)th element is assumed to lie in \([l_{ij}, c_{ij} - k_{ij}] \cup [c_{ij} + k_{ij}, r_{ij}]\). In a similar manner to Theorem 4.1, we derive the Riemannian gradients of the constraints.

We will compare the following three modelings with the above RNLO modeling. The first one is the Euclidean version of the prediction error method with prior knowledge (Euclidean nonlinear optimization; ENLO):

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{N} \| X' - (I + hA)X \|_F^2 \\
\text{subject to} & \quad l_{ij} \leq e_i^\top A e_j \leq r_{ij}, \quad ((i, j) \in I_1 \cup I_2), \\
& \quad k_{ij}^2 \leq (e_i^\top A e_j - c_{ij})^2, \quad ((i, j) \in I_2),
\end{align*}
\]  

(26a)

subject to

\[
\begin{align*}
l_{ij} \leq e_i^\top A e_j \leq r_{ij}, \quad ((i, j) \in I_1 \cup I_2), \\
k_{ij}^2 \leq (e_i^\top A e_j - c_{ij})^2, \quad ((i, j) \in I_2),
\end{align*}
\]  

(26b)

(26c)

which is obtained by replacing \((J - R)Q\) with \(A\) in (25), respectively. Since this problem does not impose the stability condition on \(A\), the solutions are not necessarily stable. The second is the Euclidean modeling that minimizes (26a) under the absence of any constraints, and the third is the Riemannian one that minimizes (25a) without the constraints.

In the experiments, we consider the case \( n = 10, h = 0.02, \) and \( N = 40 \). Indices of \( I_1 \) are randomly picked up from \( \{1, 2, 3, \ldots, 10\}^2 \) and then those of \( I_2 \) are from \( \{1, 2, 3, \ldots, 10\}^2 \setminus I_1 \). We randomly generate \((J', R', Q') \in \mathcal{M}\) using Manopt command M.rand() and then, let the true system \( A' := (J' - R')Q' \in \mathbb{R}^{n \times n} \). For each \((i, j) \in I_1, l_{ij} \) and \( r_{ij} \) are randomly generated so that \( e_i^\top A' e_j \in [l_{ij}, r_{ij}] \). As well, for each \((i, j) \in I_2, l_{ij}, r_{ij}, k_{ij}, \) and \( c_{ij} \) are randomly generated so that \( e_i^\top A' e_j \in [l_{ij}, c_{ij} - k_{ij}] \cup [c_{ij} + k_{ij}, r_{ij}] \). We set the ratios of one-box and two-box constraints as 0.2 and 0.1, respectively. Each component of \( x_0 \) is generated uniformly at random in the range \((-1000, 1000)\), and \(\{x_k\}_{k=1}^N\) is determined according to \(x_{k+1} = \exp(A'h)x_k\). Additive white Gaussian noise with a signal-to-noise ratio (SNR) of 10 dB or 20 dB is then added to these values. We also scale the data by dividing all the elements by \(\|x_0\|\) for the sake of the numerical stability. We adopt randomly generated \((J, R, Q) \in \mathcal{M}\) as the initial point.

\[ g_{ij}^1(\theta) := -e_i^\top (J - R)Qe_j + l_{ij}, \quad ((i, j) \in I_1 \cup I_2), \]

\[ g_{ij}^2(\theta) := -e_i^\top (J - R)Qe_j - c_{ij}^2 + k_{ij}^2, \quad ((i, j) \in I_2), \]

be the constraints in (25b) and (25c). The max violation at \(\theta\) is defined as

\[ \max \left\{ 0, \max_{(i, j) \in I_2} \{g_{ij}^2(\theta)\}, \max_{(i, j) \in I_1 \cup I_2} \{g_{ij}^1(\theta), g_{ij}^2(\theta)\} \right\}. \]

Throughout the experiments, we employ the identity mapping as \(B^k\) in (20), which clearly satisfies Assumption 1.

\section*{C. Numerical Results and Discussion}

We solved 50 problem instances by the six pairs of modelings and algorithms. All the algorithms were run from the identical initial point, which was produced for each problem instance, and terminated after 100 s of CPU time. The results are depicted in Figs. 1–4. Note that the scales in the subfigures within each figure significantly differ, and the result shown in Fig. 1 was obtained by randomly selecting a solution from one of the 50 problem instances. Although SNR was 10 dB for Fig. 1, it was 20 dB for the other figures. Figs. 1 and 2 demonstrate the eigenvalue distributions obtained by the various methods. Fig. 3 denotes the boxplots of the relative errors of the first largest real part of eigenvalue between true system and numerical solutions among 50 trials, respectively. Fig. 4 denotes the boxplots of the max violation among 50 trials, respectively.

As shown in Figs. 1–3, RNLO-by-RSQO was considerably better than the other methods in terms of the identification of the eigenvalues. In particular, Fig. 3 showed that only RNLO-by-RSQO can approximately estimate the most important eigenvalue, i.e., the eigenvalue with

\[ \frac{\text{SNR}}{\text{EE}} \]

\[ \frac{\text{SNR}}{\text{EEP}} \]

\[ \frac{\text{SNR}}{\text{LS}} \]

\[ \frac{\text{SNR}}{\text{GD}} \]

\[ \frac{\text{SNR}}{\text{RTR}} \]
Fig. 2. Eigenvalue distribution (SNR = 20 dB).

Fig. 3. Relative error of the first largest real part of the eigenvalue among 50 trials (SNR = 20 dB).

Fig. 4. Max violation among 50 trials (SNR = 20 dB).

the first largest real part. Here, we omitted the relative error attained by EE, because the error was considerably larger than those attained by the other methods.

While we have not shown the figure due to page limitations, an important observation needs to be highlighted. Using the same initial point, EE’s objective value was superior to that of RNLO-by-RSQO. However, in terms of eigenvalues, RNLO-by-RSQO considerably outperformed EE. Furthermore, when EE was initialized with the matrix $A$ obtained from RNLO-by-RSQO, its resulting eigenvalues were inferior to those generated by RNLO-by-RSQO. These findings emphasize EE’s tendency to overfit noisy data, contrasting with the robustness demonstrated by RNLO-by-RSQO.

According to Fig. 4, RNLO-by-RSQO was significantly better than the other methods except for EE in terms of the max violation. This result can be expected, because only RNLO-by-RSQO and EE incorporate prior information in the form of equality and inequality constraints on the parameters. The abundance of outliers implies that EE is susceptible to noise. Note that the max violation of EEP, which projects the unstable eigenvalues obtained by EE onto the imaginary axis, is larger than that of EE. This is because the projection does not take into account equality or inequality constraints.

It is remarkable that RNLO-by-RSQO is superior to GD and RTR in terms of the eigenvalues, as shown in Figs. 1–3. This results from the fact that GD and RTR consider to solve (25a) without the constraints unlike RNLO-by-RSQO. In other words, the problem for GD and RTR does not incorporate any prior knowledge other than stability unlike RNLO-by-RSQO. Thus, GD and RTR generated convergence sequences to some bad solutions as in the results of Figs. 1–3.
In summary, the superiority of RNLO-by-RSQO can be attributed to the parametrization based on Proposition 2.1 and the incorporation of prior knowledge concerning equality and inequality constraints. A particularly noteworthy point is that RNLO-by-RSQO is more robust to noise compared to EE. The only difference between these methods is whether or not they utilize the parametrization based on Proposition 2.1.

VI. CONCLUSION

We developed a prediction error method that ensures the stability of a linear system and meets with the prior knowledge. The method employs RNLO, a class of the nonlinear optimization on a Riemannian manifold. For solving this RNLO, we proposed to use RSQO. Numerical experiments with comparisons declare the effectiveness of our RNLO formulation in terms of the stability of the system and prior knowledge together with the superiority of RSQO.

An initial point generally affects numerical results in nonlinear optimization, and a sampling time \( h \) also does the accuracy of the system identification. Hence, a good proposal for the selection of an initial point and a sampling time is a possible future research. Moreover, it should be noted that our approach presented in this article does not always identify a stable system with prior knowledge information and sufficient accuracy. In fact, e.g., in situations where there is little data and a large amount of noise, the estimated matrix using our approach can be a stable matrix near the imaginary axis on the complex plane even if the true eigenvalues are far away from the axis. This issue should be resolved in the future research.

REFERENCES

[1] L. Ljung, System Identification: Theory for the User. Hoboken, NJ, USA: Prentice Hall, 2nd ed., 1999.
[2] K. Sato, H. Sato, and T. Damm, “Riemannian optimal identification method for linear systems with symmetric positive-definite matrix,” IEEE Trans. Autom. Control, vol. 65, no. 11, pp. 4493–4508, 2020.
[3] H. Sato and K. Sato, “Riemannian optimal system identification algorithm for linear MIMO systems,” IEEE Contr. Syst. Lett., vol. 1, no. 2, pp. 376–381, 2017.
[4] T. McKelvey, A. Helmersson, and T. Ribarits, “Data driven local coordinates for multivariable linear systems and their application to system identification,” Automatica, vol. 40, pp. 1629–1635, 2004.
[5] A. Wills and B. Ninness, “On gradient-based search for multivariable system estimates,” IEEE Trans. Autom. Control, vol. 53, no. 1, pp. 298–306, 2008.
[6] T. Katayama, Subspace Methods for System Identification. London, U.K.: Springer-Verlag, 2005.
[7] M. C. Campi, T. Sugie, and F. Sakai, “An iterative identification method for linear continuous-time systems,” IEEE Trans. Autom. Control, vol. 53, no. 7, pp. 1661–1669, 2008.
[8] P. V. Overschee and B. D. Moor, “N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems,” Automatica, vol. 30, no. 1, pp. 75–93, 1994.
[9] M. Verhaegen and A. Hansson, “N2SID: Nuclear norm subspace identification of innovation models,” Automatica, vol. 72, pp. 57–63, 2016.
[10] M. Verhaegen, “Identification of the deterministic part of MIMO state space models given in innovation s form from input–output data,” Automatica, vol. 30, no. 1, pp. 61–74, 1994.
[11] J. N. Kutz, S. L. Brunton, B. W. Brunton, and J. L. Proctor, Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems. Philadelphia, PA, USA: Soc. Ind. Appl. Math., 2016.
[12] J. Heiland and B. Unger, “Identification of linear time-invariant systems with dynamic mode decomposition,” Mathematics, vol. 10, no. 3, 2022, Art. no. 418.
[13] S. T. M. Dawson, M. S. Hemati, M. O. Williams, and C. W. Rowley, “Characterizing and correcting for the effect of sensor noise in the dynamic mode decomposition,” Exp. Fluids, vol. 57, 2016, Art. no. 42.
[14] T. V. Gestel, J. A. K. Suykens, P. V. Dooren, and B. D. Moor, “Identification of stable models in subspace identification by using regularization,” IEEE Trans. Autom. Control, vol. 46, no. 9, pp. 1416–1420, 2001.
[15] S. L. Lacy and D. S. Bernstein, “Subspace identification with guaranteed stability using constrained optimization,” IEEE Trans. Autom. Control, vol. 48, no. 7, pp. 1259–1263, 2003.
[16] D. N. Miller and R. A. D. Callafon, “Subspace identification with eigenvalue constraints,” Automatica, vol. 49, pp. 2468–2473, 2013.
[17] S. Siddiqi, B. Boots, and G. J. Gordon, “A constraint generation approach to learning stable linear dynamical systems,” in Proc. Adv. Neural Inf. Process. Syst., pp. 1329–1336, 2007.
[18] J. Umenberger, J. Wågberg, I. R. Manchester, and T. B. Schön, “Maximum likelihood identification of stable linear dynamical systems,” Automatica, vol. 96, pp. 280–292, 2018.
[19] J. Umenberger and R. Manchester, “Convex bounds for equation error in stable nonlinear identification,” IEEE Contr. Syst. Lett., vol. 3, no. 1, pp. 73–78, 2019.
[20] J. Umenberger and R. Manchester, “Specialized interior-point algorithm for stable nonlinear system identification,” IEEE Trans. Autom. Control, vol. 64, no. 6, pp. 2442–2456, 2019.
[21] S. Prajna, A. van der Schaft, and G. Meinsma, “An LMI approach to stabilization of port-controlled Hamiltonian systems,” Syst. Control Lett., vol. 45, pp. 371–385, 2002.
[22] K. Sato, “Riemannian optimal model reduction of stable linear systems,” IEEE Access, vol. 7, pp. 14689–14698, 2019.
[23] W. H. Yang, L.-H. Zhang, and R. Song, “Optimality conditions for the nonlinear programming problems on Riemannian manifolds,” Pacific J. Optim., vol. 10, no. 2, pp. 415–434, 2014.
[24] C. Liu and N. Boumal, “Simple algorithms for optimization on Riemannian manifolds with constraints,” Appl. Math. Optim., vol. 82, pp. 949–981, 2020.
[25] M. Obara, T. Okuno, and A. Takeda, “Sequential quadratic optimization for nonlinear optimization problems on Riemannian manifolds,” SIAM J. Optim., vol. 52, no. 3, pp. 822–853, 2022.
[26] B. Jeuris, R. Vandebril, and B. Vandereycken, “A survey and comparison of contemporary algorithms for computing the matrix geometric mean,” Electron. Trans. Numer. Anal., vol. 39, pp. 379–402, 2012.
[27] P.-A. Absil, R. Mahony, and R. Sepulchre, Optimization Algorithms on Matrix Manifolds. Princeton, NJ, USA: Princeton Univ. Press, 2008.
[28] R. Bergmann and R. Herzog, “Intrinsic formulation of KKT conditions and constraint qualifications on smooth manifolds,” SIAM J. Optim., vol. 29, no. 4, pp. 2423–2444, 2019.
[29] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, “Manopt, A matlab toolbox for optimization on manifolds,” J. Mach. Learn. Res., vol. 15, no. 1, pp. 1455–1459, 2014.
[30] V. Noferini and F. Poloni, “Nearest ω-stable matrix via Riemannian optimization,” Numer. Mathematik, vol. 148, no. 4, pp. 817–851, 2021.