

# On Input Design for Regularized LTI System Identification: Power-constrained Input <sup>★</sup>

Biqiang Mu <sup>a</sup> and Tianshi Chen <sup>b</sup>

<sup>a</sup>*Division of Automatic Control, Department of Electrical Engineering, Linköping University, Linköping, 58183, Sweden*

<sup>b</sup>*School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, 518172, China*

---

## Abstract

Input design is an important issue for classical system identification methods but has not been investigated for the kernel-based regularization method (KRM) until very recently. In this paper, we consider in the time domain the input design problem of KRMs for LTI system identification. Different from the recent result, we adopt a Bayesian perspective and in particular make use of scalar measures (e.g., the *A*-optimality, *D*-optimality, and *E*-optimality) of the Bayesian mean square error matrix as the design criteria subject to power-constraint on the input. Instead to solve the optimization problem directly, we propose a two-step procedure. In the first step, by making suitable assumptions on the unknown input, we construct a quadratic map (transformation) of the input such that the transformed input design problems are convex, the number of optimization variables is independent of the number of input data, and their global minima can be found efficiently by applying well-developed convex optimization software packages. In the second step, we derive the expression of the optimal input based on the global minima found in the first step by solving the inverse image of the quadratic map. In addition, we derive analytic results for some special types of fixed kernels, which provide insights on the input design and also its dependence on the kernel structure.

*Key words:* Input design, Bayesian mean square error, kernel-based regularization, LTI system identification, convex optimization.

---

## 1 Introduction

Over the past few years, the kernel-based regularization method (KRM), which was first introduced in Pillonetto & De Nicolao (2010) and then further developed in Chen et al. (2014, 2012); Pillonetto et al. (2011), has received increasing attention in the system identification community, see e.g., Chiuso (2016); Pillonetto et al. (2014) and the references therein. It has become a complement to the classical maximum likelihood/prediction error methods (ML/PEM), Ljung (1999); Söderström & Stoica (1989), which can be justified in a couple of aspects. First, the kernel, through

which the regularization is defined, provides a carrier for prior knowledge on the dynamic system to be identified. Second, the model complexity is tuned in a continuous manner through the hyperparameter, which is the parameter vector used to parameterize the kernel, Mu et al. (2017); Pillonetto & Chiuso (2015). Third, extensive simulation results show that KRM can have better average accuracy and robustness than ML/PEM for the data that is short and/or has low signal-to-noise ratio, Chen et al. (2012); Pillonetto et al. (2014), and as a result, algorithms of KRM have been added to the System Identification Toolbox of MATLAB (Ljung et al., 2015).

Most of the recent progress for KRM focus on the issues of kernel design and hyperparameter estimation. For the former issue, many kernels have been proposed and analyzed to embed various kinds of prior knowledge, Carli et al. (2017); Chen et al. (2016); Chen & Ljung (2017); Marconato et al. (2016); Pillonetto et al. (2016); Zorzi & Chiuso (2017). In particular, two systematic ways are introduced to design kernels in Chen & Ljung (2017): one is from a machine learning perspective which treats the impulse response as a function, and the other one is from a system theory perspective which asso-

---

<sup>★</sup> This work is supported by the Thousand Youth Talents Plan funded by the central government of China, the Shenzhen Projects Ji-20170189 and Ji-20160207 funded by the Shenzhen Science and Technology Innovation Council, the President's grant under contract No. PF. 01.000249 and the Start-up grant under contract No. 2014.0003.23 funded by the Chinese University of Hong Kong, Shenzhen, as well as by a research grant for junior researchers under contract No. 2014-5894, funded by Swedish Research Council.

*Email addresses:* [biqiang.mu@liu.se](mailto:biqiang.mu@liu.se) (Biqiang Mu), [tschen@cuhk.edu.cn](mailto:tschen@cuhk.edu.cn) (Tianshi Chen).

ciates the impulse response with a linear time-invariant (LTI) system. In Zorzi & Chiuso (2017), a harmonic analysis is first provided for existing kernels including the amplitude modulated locally stationary (AMLS) kernel introduced in Chen & Ljung (2017) and then is shown to be a useful tool to design more general kernels. In contrast, there are few results reported for the issue of hyperparameter estimation Mu et al. (2017); Pillonetto & Chiuso (2015). In particular, it was shown in Mu et al. (2017) that the Stein’s unbiased risk estimator (SURE) is asymptotically optimal but the widely used empirical Bayes estimator is not.

There are some issues for KRM that have not been addressed adequately including the issue of input design. There are numerous results on this issue for ML/PEM, see e.g., the survey papers (Gevers, 2005; Hjalmarsson, 2005; Mehra, 1974) and the books (Goodwin & Payne, 1977; Ljung, 1999; Zarrop, 1979). The current state-of-the-art of input design for ML/PEM, see e.g., Hildebrand & Gevers (2003); Hjalmarsson (2009); Jansson & Hjalmarsson (2005), is to solve the problem in a two-step procedure. The first step is to pose the problem in the frequency domain as a convex optimization problem with a linear matrix inequality constraint and then derive the optimal input power spectrum with respect to certain design criteria, and the second step is to derive the realization of the input corresponding to the optimal power spectrum. The typical design criteria are scalar measures (e.g., the trace, the determinant or the largest eigenvalue) of the *asymptotic* covariance matrix of the parameter estimate or the information matrix of ML/PEM subject to various constraints on the inputs (e.g., energy or amplitude constraints). The typical realization of the input is the filtered white noise by spectral factorization of the desired input spectrum (Hjalmarsson, 2009; Jansson & Hjalmarsson, 2005) or a multisine signal (Hildebrand & Gevers, 2003). In contrast, there have been no results reported on this issue for KRM until very recently in Fujimoto & Sugie (2016), where for a fixed kernel (a kernel with fixed hyperparameter), the optimal input is derived by maximizing the mutual information between the output and the impulse response subject to energy-constraint on the input. The proposed method in Fujimoto & Sugie (2016) is very interesting but the number of the optimization variables induced by the input design problem is equal to the number of data and thus is expensive to solve when the number of data is large. Moreover, the induced optimization problem is nonconvex and the proposed gradient-based algorithm may be inefficient and subject to local minima issue. Nevertheless, their simulation result looks quite promising and motivates the interest of further investigation.

In this paper, we treat the input design problem for KRM from a perspective different from Fujimoto & Sugie (2016). Similar to Fujimoto & Sugie (2016), we also assume that the kernel is fixed (otherwise, it can be

estimated from a preliminary experiment), but our starting point is different and is the mean square error (MSE) matrix of the regularized finite impulse response (FIR) estimate Chen et al. (2012). Since the MSE matrix depends on the unknown true impulse response, we propose to make use of the Bayesian interpretation of the KRM and derive the so-called Bayesian MSE matrix, which only depends on the fixed kernel and the input. It is then possible to use scalar measures of the Bayesian MSE matrix as the design criteria to optimize the input, e.g., the *A*-optimality, *D*-optimality, and *E*-optimality measures. Interestingly, the design criterion in Fujimoto & Sugie (2016) is equivalent to the *D*-optimality of the Bayesian MSE matrix introduced here. Moreover, we also treat the unknown input in a way different from Fujimoto & Sugie (2016) and consider power-constrained input Goodwin & Payne (1977) accordingly.

Instead to solve the optimization problem induced by the input design in a direct way, we propose a two-step procedure. In the first step, under the assumption on the unknown input, we construct a quadratic map (transformation) of the input such that the transformed input design problems are convex, the number of optimization variables is equal to the order of the FIR model, and their global minima can be found efficiently by applying well-developed convex optimization software packages, such as CVX (Grant & Boyd, 2016). In the second step, we derive the expression of the optimal input based on the global minima found in the first step by solving the inverse image of the quadratic map. From an optimization point of view, a similar optimization problem to the underlying one of our method appeared before in Hildebrand & Gevers (2003); Jansson & Hjalmarsson (2005). However, our method is essentially different. First, our input design problem is posed and solved in the time domain. Second, our method is *not* based on *asymptotic theory* and works for any finite number of data. Third, the optimal input is derived by solving the inverse image of the quadratic map. In addition, we derive analytic results for some special types of fixed kernels, which provide insights on input design and also its dependence on the kernel structure. In particular, we show that the impulsive input is globally optimal and the white noise input is asymptotically globally optimal for all diagonal kernels, but they are often not optimal for kernels with correlation, e.g., the diagonal correlated (DC) kernel introduced in Chen et al. (2012). Finally, numerical simulation results are provided to illustrate the efficacy of our method.

The remaining parts of this paper are organized as follows. In Section 2, we first review briefly KRM. In Section 3, we state the problem formulation of the input design problem. In Section 4, we introduce the two-step procedure to solve the input design problem. Then we show in Section 5 that for some fixed kernels, it is possible to derive the explicit solution of the optimal in-

put. The numerical simulation is given in Section 6 to demonstrate the proposed method. Finally, we conclude the paper in Section 7. All proofs of Theorems, Propositions, and Lemmas are postponed to the Appendix.

## 2 Regularized FIR Model Estimation

Consider a single-input single-output linear stable and causal discrete-time system

$$y(t) = G_0(q^{-1})u(t) + v(t). \quad (1)$$

Here  $t \in \mathbb{N}$  is the time index,  $q^{-1}$  is the backshift operator ( $q^{-1}u(t) = u(t-1)$ ),  $y(t), u(t) \in \mathbb{R}$  are the output and input of the system at time  $t$ , respectively,  $v(t) \in \mathbb{R}$  is a zero mean white noise with variance  $\sigma^2 > 0$  and is independent of the input  $u(t)$ , and the transfer function  $G_0(q^{-1})$  of the “true” system is

$$G_0(q^{-1}) = \sum_{k=1}^{\infty} g_k^0 q^{-k}, \quad (2)$$

where the coefficients  $g_k^0 \in \mathbb{R}, k = 1, \dots, \infty$  form the impulse response of the system. Further, we assume that the input  $u(t)$  is known (deterministic). The system identification problem is to estimate a model of  $G_0(q^{-1})$  as well as possible based on the data  $\{u(t-1), y(t)\}_{t=1}^N$ .

The stability of  $G_0(q^{-1})$  implies that its impulse response decays to zero, and thus it is always possible to truncate the infinite impulse response by a high order finite impulse response (FIR) model:

$$G(q^{-1}) = \sum_{k=1}^n g_k q^{-k}, \quad \theta = [g_1, \dots, g_n]^T \in \mathbb{R}^n, \quad (3)$$

where  $n$  is the order of the FIR model and  $(\cdot)^T$  denotes the transpose of a matrix or vector. With the FIR model (3), system (1) is written as

$$y(t) = \phi^T(t)\theta + v(t), \quad \phi(t) = [u(t-1), \dots, u(t-n)]^T, \quad (4)$$

and its matrix-vector form is

$$Y = \Phi\theta + V, \quad (5)$$

$$Y = [y(1) \ y(2) \ \dots \ y(N)]^T, \quad (6)$$

$$\Phi = [\phi(1) \ \phi(2) \ \dots \ \phi(N)]^T, \quad (7)$$

$$V = [v(1) \ v(2) \ \dots \ v(N)]^T, \quad (8)$$

where the unknown input  $u(t)$  with  $t = -1, \dots, -n+1$ , can be handled in different ways: e.g., they can be not used (“non-windowed”) or they can be set to zero (“pre-windowed”); see (Ljung, 1999, p. 320) for discussions.

There are different methods to estimate  $\theta$  and the simplest one is perhaps the least squares (LS) method:

$$\hat{\theta}_N^{\text{LS}} = \arg \min_{\theta} \|Y - \Phi\theta\|^2 = (\Phi^T \Phi)^{-1} \Phi^T Y, \quad (9)$$

where  $\|\cdot\|$  is the Euclidean norm. However,  $\hat{\theta}_N^{\text{LS}}$  may have large variance and thus large mean square error (MSE), as its variance increases approximately linearly with respect to  $n$ . One way to mitigate the possible large variance and reduce the MSE is by using the regularized least squares (RLS) method, see e.g., Chen et al. (2012):

$$\hat{\theta}_N^{\text{R}} = \arg \min_{\theta \in \mathbb{R}^n} \|Y - \Phi\theta\|^2 + \sigma^2 \theta^T P^{-1} \theta \quad (10a)$$

$$= P\Phi^T(\Phi P\Phi^T + \sigma^2 I_N)^{-1} Y, \quad (10b)$$

where  $P$  is positive semidefinite and is called the kernel matrix ( $\sigma^2 P^{-1}$  is often called the regularization matrix), and  $I_N$  is the  $N$ -dimensional identity matrix.

Now we let  $\theta_0 = [g_1^0, \dots, g_n^0]^T$ , where  $g_1^0, \dots, g_n^0$  are defined in (2) and we then obtain the MSE matrix of  $\hat{\theta}_N^{\text{R}}$ :

$$\hat{\theta}_N^{\text{R}} - \theta_0 = -\sigma^2 Q^{-1} P^{-1} \theta_0 + Q^{-1} \Phi^T V, \quad (11)$$

$$\begin{aligned} M_N &= E(\hat{\theta}_N^{\text{R}} - \theta_0)(\hat{\theta}_N^{\text{R}} - \theta_0)^T \\ &= \sigma^4 Q^{-1} P^{-1} \theta_0 \theta_0^T P^{-1} Q^{-1} + \sigma^2 Q^{-1} \Phi^T \Phi Q^{-1} \end{aligned} \quad (12)$$

$$Q = \Phi^T \Phi + \sigma^2 P^{-1}, \quad (13)$$

where  $E(\cdot)$  is the mathematical expectation. It has been shown in Mu et al. (2017, Prop. 2) that for a suitably chosen kernel matrix  $P$ ,  $\text{Tr}(M_N(\hat{\theta}_N^{\text{R}})) \leq \text{Tr}(M_N(\hat{\theta}_N^{\text{LS}}))$ , where  $\text{Tr}(\cdot)$  is the trace of a square matrix.

The problem to achieve a good  $\hat{\theta}_N^{\text{R}}$  boils down to the choice of a suitable kernel matrix  $P$ , which contains two issues: kernel design and hyperparameter estimation.

### 2.1 Kernel Design

The issue of kernel design is regarding how to embed in a kernel the prior knowledge of the underlying system to be identified by parameterizing the kernel with a parameter vector, say  $\eta$ , called hyperparameter. The essence of kernel design is analogous to the model structure design for ML/PEM, and the kernel determines the underlying model structure for the regularized FIR model (10b). So far, several kernels have been proposed, such as the stable spline (SS) kernel (Pillonetto & De Nicolao, 2010), the diagonal correlated (DC) kernel and the tuned-correlated (TC) kernel (Chen et al., 2012), the latter two of which are defined as follows:

$$\text{DC} : P_{kj}(\eta) = c\lambda^{(k+j)/2} \rho^{|j-k|},$$

$$\eta = [c, \lambda, \rho] \in \Omega = \{c \geq 0, 0 \leq \lambda \leq 1, |\rho| \leq 1\}; \quad (14)$$

$$\begin{aligned} \text{TC : } P_{kj}(\eta) &= c\lambda^{\max(k,j)}, \\ \eta &= [c, \lambda] \in \Omega = \{c \geq 0, 0 \leq \lambda \leq 1\}. \end{aligned} \quad (15)$$

where the TC kernel (15) is a special case of the DC kernel with  $\rho = \sqrt{\lambda}$  (Chen et al., 2012) and is also called the first order SS kernel (Pillonetto et al., 2014).

## 2.2 Hyperparameter Estimation

Once a kernel is designed, the next step is to determine the hyperparameter based on the data. The essence of hyperparameter estimation is analogous to the model order selection for ML/PEM, and the hyperparameter determines the model complexity of the regularized FIR model (10b). Several estimation methods have been suggested in (Pillonetto et al., 2014, Section 14). The most widely used method is the empirical Bayes (also called marginal likelihood maximization) method. The idea is to adopt the Bayesian perspective and embed the regularization term  $\sigma^2 \theta^T P^{-1} \theta$  in a Bayesian framework. More specifically, we assume in (4) that  $v(t)$  and  $\theta$  are independent and Gaussian distributed with  $v(t) \sim \mathcal{N}(0, \sigma^2)$  and

$$\theta \sim \mathcal{N}(0, P). \quad (16)$$

Then  $\theta$  and  $Y$  are jointly Gaussian distributed and moreover, the posterior distribution of  $\theta$  given  $Y$  is

$$\begin{aligned} \theta|Y &\sim \mathcal{N}(\hat{\theta}^R, \hat{P}^R), \\ \hat{\theta}_N^R &= P\Phi^T(\Phi P\Phi^T + \sigma^2 I_N)^{-1}Y, \\ \hat{P}_N^R &= P - P\Phi^T(\Phi P\Phi^T + \sigma^2 I_N)^{-1}\Phi P = \sigma^2 Q^{-1}. \end{aligned} \quad (17)$$

Moreover, the output  $Y$  is also Gaussian with zero mean and covariance matrix  $\Phi P(\eta)\Phi^T + \sigma^2 I_N$ . Therefore, the hyperparameter  $\eta$  can be estimated by maximizing the marginal likelihood, or equivalently,

$$\begin{aligned} \text{EB: } \hat{\eta} &= \arg \min_{\eta \in \Omega} Y^T F(\eta)^{-1} Y + \log \det(F(\eta)), \quad (18) \\ F &= \Phi^T P \Phi + \sigma^2 I_N, \end{aligned} \quad (19)$$

where  $\det(\cdot)$  is the determinant of a matrix. The EB (18) has the advantage that it is robust Pillonetto & Chiuso (2015), but not asymptotical optimal in the sense of MSE (Mu et al., 2017).

## 3 Problem Formulation for Input Design

### 3.1 Bayesian MSE Matrix and Design Criteria

The goal of input design is to determine, under suitable conditions, an input sequence such that the regularized FIR model (10b) is as good as possible. The MSE matrix

(12) is a natural measure to evaluate how good the regularized FIR model estimate (10b) is. Unfortunately, the first term of (12) depends on the true impulse response  $\theta_0$  and thus it can not be used directly.

There are different ways to deal with this difficulty. We adopt a Bayesian perspective that is similar to the derivation of (18). More specifically, we assume that

$$\theta_0 \sim \mathcal{N}(0, P). \quad (20)$$

Then taking expectation on both side of (12) leads to

$$\begin{aligned} M_N &= \sigma^4 Q^{-1} P^{-1} Q^{-1} + \sigma^2 Q^{-1} \Phi^T \Phi Q^{-1} \\ &= \sigma^2 Q^{-1} (\sigma^2 P^{-1} + \Phi^T \Phi) Q^{-1} \\ &= \sigma^2 Q^{-1}, \text{ "Bayesian MSE"}. \end{aligned} \quad (21)$$

Since the Bayesian perspective is adopted, (21) is called the Bayesian MSE matrix of the regularized FIR model estimate (10b) under the assumption (20). Interestingly, the Bayesian MSE matrix (21) is equal to the posterior covariance of the regularized FIR model estimate (17) under the assumption (16).

We will tackle the input design problem by minimizing a scalar measure of the Bayesian MSE matrix (21) of the regularized FIR model estimate (10b). This idea is similar to that of the traditional input design problem by minimizing a scalar measure of the asymptotic covariance matrix of the parameter estimate (Ljung, 1999). The typical  $A$ -optimality,  $D$ -optimality, and  $E$ -optimality scalar measures of the Bayesian MSE matrix (21) will be chosen as the design criteria for the input design problem and given below:

$$D - \text{optimality : } \det(M_N) \triangleq \det(\sigma^2 Q^{-1}), \quad (22)$$

$$A - \text{optimality : } \text{Tr}(M_N) \triangleq \text{Tr}(\sigma^2 Q^{-1}), \quad (23)$$

$$E - \text{optimality : } \lambda_{\max}(M_N) \triangleq \lambda_{\max}(\sigma^2 Q^{-1}), \quad (24)$$

where  $\lambda_{\max}(\cdot)$  is the largest eigenvalue of a matrix.

**Remark 1** Another possible way to deal with the unknown  $\theta_0$  in the MSE matrix (12) is to replace  $\theta_0$  by an estimate obtained in advance, for example, the LS estimate (9). In this case, the MSE matrix (12) becomes

$$\sigma^4 Q^{-1} P^{-1} \hat{\theta}_N^{\text{LS}} (\hat{\theta}_N^{\text{LS}})^T P^{-1} Q^{-1} + \sigma^2 R^{-1} \Phi^T \Phi Q^{-1}. \quad (25)$$

Clearly, the input design would depend on the quality of the chosen estimate.

**Remark 2** Interestingly, the  $D$ -optimality measure (22) is equivalent to the mutual information between the true impulse response and the output  $Y$  given in Fujimoto & Sugie (2016).

### 3.2 Problem Statement

Now we are able to state the input design problem based on the design criteria (22)–(24).

Assume that the kernel matrix  $P(\eta)$  and the variance  $\sigma^2$  of the noise are known in advance (otherwise, they can be estimated with a preliminary experiment) and also note that the construction of the regression matrix  $\Phi$  in (7) requires the inputs  $u_{-n+1}, \dots, u_{N-1}$ . Then the input design problem is to determine an input sequence

$$u_{-n+1}, \dots, u_{-1}, u_0, u_1, \dots, u_{N-1}, \quad (26)$$

where, for simplicity,  $u_t$  is used to denote  $u(t)$  hereafter, such that the input sequence (26) minimizes the design criteria (22)–(24) subject to certain constraints.

It should be noted that the inputs  $u_{-n+1}, \dots, u_{-1}$  are unknown for the identification problem but can be treated as design variables for the input design problem. In particular, Fujimoto & Sugie (2016) sets the inputs  $u_{-n+1}, \dots, u_{-1}$  to be zero, i.e.,

$$u_{-n+1} = \dots = u_{-1} = 0, \quad (27)$$

and essentially minimizes the  $D$ -optimality measure (22) subject to an energy-constraint

$$\begin{aligned} \mathcal{U} &= \{u | u^T u \leq \mathcal{E}, u \in \mathbb{R}^N\}, \\ u &= [u_0, u_1, \dots, u_{N-1}]^T, \end{aligned} \quad (28)$$

where  $\mathcal{E} > 0$  is a known constant and is the maximum available energy for the input.

Here, we treat  $u_{-n+1}, \dots, u_{-1}$  in a different way. Our idea is not only to reduce the number of optimization variables of the input design problem, but also to bring it certain structure such that it becomes easier to solve, see Section 4 for details. In particular, we assume  $N \geq n$  and set

$$u_{-i} = u_{N-i}, \quad i = 1, \dots, n-1. \quad (29)$$

The advantage of doing so is that the regression matrix

$\Phi$  has a good structure and becomes a circulant matrix:

$$\Phi = \begin{bmatrix} u_0 & u_{N-1} & \cdots & u_{N-n+2} & u_{N-n+1} \\ u_1 & u_0 & \cdots & u_{N-n+3} & u_{N-n+2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ u_{n-2} & u_{n-3} & \cdots & u_0 & u_{N-1} \\ u_{n-1} & u_{n-2} & \cdots & u_1 & u_0 \\ u_n & u_{n-1} & \cdots & u_2 & u_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ u_{N-1} & u_{N-2} & \cdots & u_{N-n+1} & u_{N-n} \end{bmatrix}.$$

Moreover, we consider the power constraint on the input sequence (26) used for the traditional ML/PEM input design problem, see e.g., (Goodwin & Payne, 1977, p. 129, eq. (6.3.12)), i.e.,

$$\sum_{t=1}^N u_{t-i}^2 = \mathcal{E}, \quad i = 1, \dots, n, \quad (30)$$

Under the assumption (29), the input design problem of minimizing the design criteria (22)–(24) subject to the constraint (30) can be equivalently written as follows:

$$u_D^* \triangleq \arg \min_{u \in \mathcal{U}} \det(\sigma^2 Q^{-1}), \quad (31)$$

$$u_A^* \triangleq \arg \min_{u \in \mathcal{U}} \text{Tr}(\sigma^2 Q^{-1}), \quad (32)$$

$$u_E^* \triangleq \arg \min_{u \in \mathcal{U}} \lambda_{\max}(\sigma^2 Q^{-1}), \quad (33)$$

where the constraint (30) under the assumption (29) becomes

$$\begin{aligned} \mathcal{U} &= \{u | u^T u \leq \mathcal{E}, u \in \mathbb{R}^N\}, \\ u &= [u_0, u_1, \dots, u_{N-1}]^T, \end{aligned} \quad (34)$$

It is worth to note that the optimal input is not unique in general. To check this, note that if an input sequence  $u^*$  is optimal, then  $-u^*$  is also optimal, where  $u^*$  can be any of  $u_D^*, u_A^*$  and  $u_E^*$ . Therefore,  $u_D^*, u_A^*, u_E^*$  should be understood as the set of all optimal inputs minimizing (31), (32), (33), respectively.

## 4 Main Results

Similar to Fujimoto & Sugie (2016), we can also try to solve the input design problems (31)–(33) by using gradient-based algorithms. However, such algorithms may have the following problems:

- 1) the input design problems (31)–(33) is expensive to solve for large  $N$ , because the number of optimization variables is  $N$ .



- 2) the gradient-based algorithms may be subject to the issue of local minima, because the input design problems (31)–(33) are non-convex.

We propose to solve the input design problems (31)–(33) in a two-step procedure, and its idea is sketched below:

- 1) under the assumption (29), we construct a quadratic map (transformation) of  $u$ , say  $f(u)$ , such that the transformed input design problems of (31)–(33) are convex and thus can be solved efficiently.
- 2) by exploiting the properties of  $f(\cdot)$  and based on the globally minima of the transformed input design problems of (31)–(33), we derive the optimal inputs by solving the inverse of the map  $f(\cdot)$ .

#### 4.1 A Quadratic Map

Under the assumption (29), we define

$$\begin{aligned} r &= [r_0, r_1, \dots, r_{n-1}]^T, \\ r_j &= \sum_{k=0}^{N-1} u_k u_{k-j}, \quad j = 0, \dots, n-1. \end{aligned} \quad (35)$$

Then we have

$$R \triangleq \Phi^T \Phi = \begin{bmatrix} r_0 & r_1 & \cdots & r_{n-2} & r_{n-1} \\ r_1 & r_0 & \ddots & r_{n-3} & r_{n-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ r_{n-2} & r_{n-3} & \ddots & r_0 & r_1 \\ r_{n-1} & r_{n-2} & \cdots & r_1 & r_0 \end{bmatrix}, \quad (36)$$

which is a positive semidefinite Toeplitz matrix.

Actually, the definition (35) determines a quadratic vector-valued function

$$r = f(u) = [f_0(u), \dots, f_{n-1}(u)]^T. \quad (37)$$

The domain of  $f(\cdot)$  is  $\mathcal{U}$  which has been defined in (34) and is convex and compact, and the  $j$ -element of  $f(\cdot)$  is

$$f_j(u) = u^T L_j u, \quad j = 0, \dots, n-1, \quad (38)$$

$$L_j = \frac{1}{2} \begin{bmatrix} 0 & I_{N-j} \\ I_j & 0 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 0 & I_j \\ I_{N-j} & 0 \end{bmatrix}. \quad (39)$$

It follows that the corresponding value  $r_0$  is  $\mathcal{E}$  for any input  $u \in \mathcal{U}$ . Moreover, denote the image of  $f(\cdot)$  under  $\mathcal{U}$  by

$$\mathcal{F} = \{f(u) | u \in \mathcal{U}\}, \quad (40)$$

which is a convex polytope (See Theorem 2 below).

#### 4.2 Transformed Input Design Problem

Interestingly, the quadratic map (37) makes the transformed input design problems of (31) to (33) convex.

To show this, note that the matrix  $Q$  in (13) can be written as

$$Q(r) = r_0 I_n + r_1 Q_1 + \cdots + r_{n-1} Q_{n-1} + \sigma^2 P^{-1}, \quad (41)$$

where  $Q_i \in \mathbb{R}^{n \times n}$ ,  $i = 1, \dots, n-1$  is symmetric and zero everywhere except the  $(jl)$ -element which is one if  $|j-l| = i$  with  $j, l = 1, \dots, n$ . The map (37) transforms the input design problems (31)–(33) to the following optimization problems, respectively,

$$r_D^* = \arg \min_{r \in \mathcal{F}} \log \det(\sigma^2 Q(r)^{-1}) \quad (42)$$

$$r_A^* = \arg \min_{r \in \mathcal{F}} \text{Tr}(\sigma^2 Q(r)^{-1}) \quad (43)$$

$$r_E^* = \arg \min_{r \in \mathcal{F}} \lambda_{\max}(\sigma^2 Q(r)^{-1}) \quad (44)$$

where  $\log \det(\cdot)$  is used instead of  $\det(\cdot)$ .

**Proposition 1** Consider the transformed input design problems (42) to (44). It follows that

- The problems (42) and (43) are strictly convex and thus  $r_D^*$  and  $r_A^*$  are unique in  $\mathcal{F}$ .
- The problem (44) is convex and  $r_E^*$  exists but may be not unique in  $\mathcal{F}$ .

**Remark 3** From an optimization point of view, the optimization problems (42) to (44), modulo the constraint, appeared before, see e.g., Jansson & Hjalmarsson (2005). However, the contexts are different: the elements of  $r$  are the auto-correlation coefficients of the input power spectrum in Jansson & Hjalmarsson (2005) but they do not have such interpretation here unless we divide them by  $N$  and let  $N$  go to  $\infty$ . In addition, the positivity constraint of the input spectrum is transformed into a linear matrix inequality by applying the Kalman-Yakubovich-Popov lemma in Jansson & Hjalmarsson (2005), while the constraint here is the feasible set  $\mathcal{F}$ , which is a convex polytope (See Remark 6 below).

**Remark 4** An optimal  $r^*$  satisfies  $r_0^* = \mathcal{E}$ , where  $r_0^*$  is the first element of  $r^*$ , and  $r^*$  could be  $r_D^*$ ,  $r_A^*$ , and  $r_E^*$ .

**Remark 5** For  $i = 1, \dots, n-1$ , the elements of the gradient of (42) and (43) are respectively given by

$$\frac{\partial \log \det(\sigma^2 Q(r)^{-1})}{\partial r_i} = -\text{Tr}(Q(r)^{-1} Q_i), \quad (45)$$

$$\frac{\partial \text{Tr}(\sigma^2 Q(r)^{-1})}{\partial r_i} = -\sigma^2 \text{Tr}(Q(r)^{-2} Q_i). \quad (46)$$

### 4.3 Finding Optimal Inputs

Now we assume that the global minima of the transformed input design problems (42) to (44) have been found and denoted by  $r^* \in \mathcal{F}$ , which could be any one of  $r_D^*$ ,  $r_A^*$ , and  $r_E^*$ . Then we consider the problem how to derive the optimal input  $u^*$  from  $r^* \in \mathcal{F}$  by exploiting the properties of the quadratic map  $f(\cdot)$  and the inverse map of  $f(\cdot)$  for a given vector  $r \in \mathcal{F}$ .

#### 4.3.1 Properties of the Quadratic Map

The map  $f(\cdot)$  has the following properties, which ease the derivation of the optimal input as will be seen shortly.

**Theorem 1** Consider the map  $f(\cdot)$  defined in (37). Then there exists a matrix  $S \in \mathbb{R}^{n \times N}$  and an orthogonal matrix  $W \in \mathbb{R}^{N \times N}$  such that the map  $f(\cdot)$  can be written in a composite form as follows:

$$f(u) = h_1(h_2(h_3(u))) \quad \text{with} \quad (47)$$

$$h_1(x) = Sx \quad (48)$$

$$h_2(z) = [z_0^2, z_1^2, \dots, z_{N-1}^2]^T \quad (49)$$

$$h_3(u) = W^T u \quad (50)$$

where  $x = [x_0, x_1, \dots, x_{N-1}]^T$ ,  $z = [z_0, z_1, \dots, z_{N-1}]^T$ . Moreover, the image of  $h_3(\cdot)$  is

$$\mathcal{Z} = \{h_3(u) | u \in \mathcal{U}\} = \{z | z^T z = \mathcal{E}\},$$

the image of  $h_2(\cdot)$

$$\mathcal{X} = \{h_2(z) | z \in \mathcal{Z}\} \quad (51)$$

$$= \left\{ x \mid \sum_{i=0}^{N-1} x_i = \mathcal{E}, x_i \geq 0, i=0, 1, \dots, N-1 \right\}, \quad (52)$$

is a convex polytope and the image of  $h_1(\cdot)$

$$\mathcal{F} = \{f(u) | u \in \mathcal{U}\} = \{Sx | x \in \mathcal{X}\} \quad (53)$$

is also a convex polytope.

One choice for the matrices  $S$  and  $W$  in Theorem 1 is given in the following theorem.

**Theorem 2** Let  $\varpi = 2\pi/N$  and define the vectors

$$\xi_j = \begin{bmatrix} 1 \\ \cos(j\varpi) \\ \cos(2j\varpi) \\ \vdots \\ \cos((N-1)j\varpi) \end{bmatrix}, \quad \zeta_j = \begin{bmatrix} 0 \\ \sin(j\varpi) \\ \sin(2j\varpi) \\ \vdots \\ \sin((N-1)j\varpi) \end{bmatrix}$$

for  $j = 0, 1, \dots$ . Then, one choice for the matrices  $W$  and  $S$  in Theorem 1 is

1) when  $N$  is even,

$$W = \sqrt{\frac{2}{N}} \begin{bmatrix} \frac{\xi_0}{\sqrt{2}}, \xi_1, \dots, \xi_{\frac{N-2}{2}}, \frac{\xi_{\frac{N}{2}}}{\sqrt{2}}, \zeta_{\frac{N-2}{2}}, \dots, \zeta_1 \end{bmatrix} \quad (54)$$

$$S = \begin{bmatrix} \xi_0, \xi_1, \dots, \xi_{n-1} \end{bmatrix}^T \\ = \begin{bmatrix} \xi_0(1:n), \xi_1(1:n), \dots, \xi_{\frac{N}{2}}(1:n), \dots, \xi_1(1:n) \end{bmatrix} \quad (55)$$

and  $\text{rank}(S) = \min(N/2+1, n)$ . Also,

$$\mathcal{F} = \text{conv} \left\{ \mathcal{E}\xi_0(1:n), \mathcal{E}\xi_1(1:n), \dots, \mathcal{E}\xi_{\frac{N}{2}}(1:n) \right\} \quad (56)$$

where  $\text{conv}\{\dots\}$  is the convex hull of the corresponding vectors and  $\xi_j(1:n)$  is the vector consisting of the first  $n$  elements of  $\xi_j$ .

2) when  $N$  is odd,

$$W = \sqrt{\frac{2}{N}} \begin{bmatrix} \frac{\xi_0}{\sqrt{2}}, \xi_1, \dots, \xi_{\frac{N-1}{2}}, \zeta_{\frac{N-1}{2}}, \dots, \zeta_1 \end{bmatrix} \quad (57)$$

$$S = \begin{bmatrix} \xi_0, \xi_1, \dots, \xi_{n-1} \end{bmatrix}^T \\ = \begin{bmatrix} \xi_0(1:n), \xi_1(1:n), \dots, \xi_{\frac{N-1}{2}}(1:n), \\ \xi_{\frac{N-1}{2}}(1:n), \dots, \xi_1(1:n) \end{bmatrix} \quad (58)$$

and  $\text{rank}(S) = \min((N+1)/2, n)$ . In addition,

$$\mathcal{F} = \text{conv} \left\{ \mathcal{E}\xi_0(1:n), \mathcal{E}\xi_1(1:n), \dots, \mathcal{E}\xi_{\frac{N-1}{2}}(1:n) \right\}. \quad (59)$$

**Remark 6** The transformed input design problems (42), (43), (44) can be solved effectively by CVX (Grant & Boyd, 2016). The optimization criteria of (42), (43), (44) are standard in CVX and we only need to rewrite the feasible set  $\mathcal{F}$  as a number of linear equalities and inequalities by the definition of convex hull (56) or (59):

• when  $N$  is even, we have from (56) that

$$\mathcal{E}S(:, 1:N/2+1)a = r \\ a = [a_1, a_2, \dots, a_{N/2+1}]^T, \quad a_j \geq 0 \quad (60)$$

where  $r \in \mathcal{F}$ ,  $S(:, 1:N/2+1)$  denotes the first  $N/2+1$  columns of  $S$ , and the first equality constraint is  $\sum_{j=1}^{N/2+1} a_j = 1$  since the first element of  $r$  is  $\mathcal{E}$  and all the elements of the first row of  $S$  are one.

- when  $N$  is odd, we have from (59) that

$$\begin{aligned} \mathcal{E}S(:, 1:(N+1)/2)a &= r \\ a &= [a_1, a_2, \dots, a_{(N+1)/2}]^T, \quad a_j \geq 0. \end{aligned} \quad (61)$$

Since the columns of  $S$  are symmetric (See (55) and (58)),  $\mathcal{F}$  can be uniformly expressed by for even and odd  $N$

$$\begin{aligned} \mathcal{E}Sa &= r \\ a &= [a_1, a_2, \dots, a_N]^T, \quad a_j \geq 0. \end{aligned} \quad (62)$$

#### 4.3.2 Derivation of the Optimal Input

Given any  $r \in \mathcal{F}$ , the inverse image  $f^{-1}(r)$  of  $f(\cdot)$  can be derived in the following way based on (47)–(50):

- 1) we find the inverse image of  $h_1(\cdot)$  for  $r \in \mathcal{F}$ :

$$\begin{aligned} \mathcal{X}(r) &\triangleq \{x | Sx = r, x \in \mathcal{X}\} \\ &= \{\mathcal{K}(S) \oplus S^\dagger r\} \cap \mathcal{X} \end{aligned} \quad (63)$$

where  $S^\dagger$  is the Moore-Penrose pseudoinverse of  $S$  and  $\mathcal{K}(S) = \{x | Sx = 0\}$  is the null space of  $S$ , and  $\oplus$  is the direct sum between two subspace.

- 2) we find the inverse image of  $h_2(\cdot)$  for  $x \in \mathcal{X}(r)$ :

$$\begin{aligned} \mathcal{Z}(r) &\triangleq \{z | h_2(z) \in \mathcal{X}(r)\} \\ &= \{[\pm\sqrt{x_0}, \dots, \pm\sqrt{x_{N-1}}]^T | x \in \mathcal{X}(r)\}. \end{aligned} \quad (64)$$

- 3) we find the inverse image of  $h_3(\cdot)$  for  $z \in \mathcal{Z}(r)$ :

$$\mathcal{U}(r) \triangleq \{u | W^T u \in \mathcal{Z}(r)\} = \{Wz | z \in \mathcal{Z}(r)\}. \quad (65)$$

It is worth to note that  $\mathcal{X}(r)$  is still a convex polytope.

**Proposition 2** *The inverse image  $\mathcal{X}(r)$  is a convex polytope and is determined by a group of halfspaces and hyperplanes:*

$$\mathcal{X}(r) = \{x | Sx = r, x_i \geq 0, i = 0, \dots, N-1\}. \quad (66)$$

This means that the optimal inputs  $u^* = \mathcal{U}(r^*)$ , where  $u^*$  can be any of  $u_D^*, u_A^*$  and  $u_E^*$ .

**Remark 7** *Noting that  $W$  is orthogonal and the row vectors of  $S$  are part of the column vectors of  $W$  subject to some scaling factors, a basis of  $\mathcal{K}(S)$  can be derived as follows:*

- when  $N$  is even, a basis of  $\mathcal{K}(S)$  is

$$\begin{cases} \{\zeta_1, \dots, \zeta_{\frac{N-2}{2}}\} & \text{if } n-1 \geq N/2, \\ \{\xi_n, \dots, \xi_{\frac{N}{2}}, \zeta_1, \dots, \zeta_{\frac{N-2}{2}}\} & \text{if } n-1 < N/2. \end{cases}$$

- when  $N$  is odd, a basis of  $\mathcal{K}(S)$  is

$$\begin{cases} \{\zeta_1, \dots, \zeta_{\frac{N-1}{2}}\} & \text{if } n-1 \geq (N-1)/2, \\ \{\xi_n, \dots, \xi_{\frac{N-1}{2}}, \zeta_1, \dots, \zeta_{\frac{N-2}{2}}\} & \text{if } n-1 < (N-1)/2. \end{cases}$$

**Remark 8** *As can be seen from the algorithm given above for finding the optimal inputs  $u^*$  from  $r^* \in \mathcal{F}$ , the key is to determine the set  $\mathcal{X}(r^*)$  while steps 2) and 3) are straightforward. Actually, some elements belonging to  $\mathcal{X}(r^*)$  can be derived from the global minima of the optimization problems (42)–(44). Assume that the global minima corresponding to the constraints (60) is  $\{r^*, a^*\}$ , respectively, where  $a^* = [a_1^*, a_2^*, \dots, a_{N/2+1}^*]^T$ . Then all vectors*

$$\mathcal{E} \left[ a_1^*, \alpha_2 a_2^*, \dots, \alpha_{\frac{N}{2}} a_{\frac{N}{2}}^*, a_{\frac{N}{2}+1}^*, (1-\alpha_{\frac{N}{2}}) a_{\frac{N}{2}}^*, \dots, (1-\alpha_2) a_2^* \right]^T$$

*belong to  $\mathcal{X}(r^*)$  for  $0 \leq \alpha_i \leq 1, i = 2, \dots, N/2$ . Accordingly, for the constraint (61) with  $a^* = [a_1^*, a_2^*, \dots, a_{\frac{N+1}{2}}^*]^T$ , we have all vectors*

$$\mathcal{E} \left[ a_1^*, \alpha_2 a_2^*, \dots, \alpha_{\frac{N+1}{2}} a_{\frac{N+1}{2}}^*, (1-\alpha_{\frac{N+1}{2}}) a_{\frac{N+1}{2}}^*, \dots, (1-\alpha_2) a_2^* \right]^T$$

*belong to  $\mathcal{X}(r^*)$  for  $0 \leq \alpha_i \leq 1, i = 2, \dots, (N+1)/2$ . At last, for the constraint (62) with  $a^* = [a_1^*, a_2^*, \dots, a_N^*]^T$ , we have the vector  $\mathcal{E}a^*$  belongs to  $\mathcal{X}(r^*)$ .*

## 5 Optimal Input for Some Special Cases

In general, there is no analytic solution to the input design problems. In this section, we study the optimal input for some special types of fixed kernels. The obtained analytic results provide insights on input design and also its dependence on the kernel structure.

### 5.1 Diagonal Kernel Matrices

We first consider the ridge kernel matrix and then more general diagonal kernel matrices.

#### 5.1.1 Ridge Kernel Matrix

When the kernel matrix is

$$P = cI_n, \text{ with } c > 0, \quad (67)$$

it is possible to obtain the explicit expression of  $r_D^*$  and  $r_A^*$ , which is given in the following proposition.



**Proposition 3** Consider the ridge kernel matrix (67). Then we have

$$r_D^* = r_A^* = [\mathcal{E}, \underbrace{0, \dots, 0}_{n-1 \text{ zeros}}]^T \triangleq r^\dagger. \quad (68)$$

Proposition 3 shows that for ridge kernel matrix (67), any input such that  $\Phi^T \Phi = \mathcal{E} I_n$  is the global minimum of (31) and (32). The optimal input given in Proposition 3 is identical with the classic result of input design for FIR model estimation with ML/PEM given in Goodwin & Payne (1977). In fact, this result is not surprising since the ridge kernel assumes that the prior distribution of the true impulse response is independent and identically distributed Gaussian, which does not provide any information on the correlation and the decay rate of the true impulse response. The optimal solution (68) implies that the optimal inputs could be:

- 1) Impulsive inputs:  $[\pm\sqrt{\mathcal{E}}, 0, \dots, 0]^T$ ;
- 2) White noise inputs as  $N \rightarrow \infty$ .

Note that the impulsive input is globally optimal since it satisfies (68) exactly. In contrast, it was only shown in Fujimoto & Sugie (2016) that the impulsive input is *locally* optimal. In addition, as  $N \rightarrow \infty$ , the white noise input is also globally optimal. Clearly, the white noise input is *approximately* globally optimal for large  $N$ .

### 5.1.2 General Diagonal Kernel Matrices

When the kernel matrix is

$$P = \text{diag}([\lambda_1, \lambda_2, \dots, \lambda_n]) \quad (69)$$

where  $\lambda_i > 0$ ,  $i = 1, \dots, n$ , and  $\text{diag}(\cdot)$  represents a diagonal matrix, the optimal  $r_D^*$  and  $r_A^*$  are same as the one in Proposition 3.

**Theorem 3** Consider the diagonal kernel matrix (69). Then we have

$$r_D^* = r_A^* = r^\dagger. \quad (70)$$

Note that the diagonal kernel matrix (69) contains the DI kernel matrix studied in Chen et al. (2012) as a special case, which takes the following form

$$P = \text{diag}(c[\lambda, \lambda^2, \dots, \lambda^n]), \quad c, \lambda > 0.$$

Theorem 3 shows that for general diagonal kernel matrices the information on the decay rate of the impulse response is not useful for the input design, if no information on the correlation of the true impulse response is provided.

## 5.2 Nondiagonal Kernel Matrices

Theorem 3 motivates an interesting question that, for nondiagonal kernel matrices, whether  $r_D^*$  and  $r_A^*$  are still equal to  $r^\dagger$  or not.

### 5.2.1 Kernel matrix with a tridiagonal inverse

To answer this question, we first consider a class of kernel matrices  $P$  such that it is nonsingular and its inverse is tridiagonal, i.e.,  $P^{-1}$  can be written as follows:

$$P^{-1} = \begin{bmatrix} p_1 & -e_2 & & & \\ -e_2 & p_2 & -e_3 & & \\ & \ddots & \ddots & \ddots & \\ & & -e_{n-1} & p_{n-1} & -e_n \\ & & & -e_n & p_n \end{bmatrix} \quad (71)$$

where  $p_i > 0$  for  $1 \leq i \leq n$ . Then, we have the following result.

**Theorem 4** Consider a kernel matrix  $P$  which has tridiagonal inverse (71). Then we have

$$r_D^* \neq r^\dagger \text{ and } r_A^* \neq r^\dagger \quad (72)$$

for the following two cases:

- 1)  $e_j > 0$  for  $j = 2, \dots, n$  with any  $N$ ;
- 2)  $e_j < 0$  for  $j = 2, \dots, n$  with even  $N$ .

One may wonder whether there exist kernel matrices such that its inverse is tridiagonal. The answer is positive. Actually, the DC kernel (14) has tridiagonal inverse Carli et al. (2017), and moreover, the simulation induced kernels with Markov property of order 1 also have tridiagonal inverse (Chen & Ljung, 2017, Section 4.5). In particular, the inverse of the DC kernel matrix is

$$P_{DC}^{-1} = \frac{1}{c(1-\rho^2)} \begin{bmatrix} \frac{1}{\lambda} & -\frac{\rho}{\lambda^{\frac{3}{2}}} & \dots & 0 & 0 \\ -\frac{\rho}{\lambda^{\frac{3}{2}}} & \frac{1+\rho^2}{\lambda^2} & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ddots & \frac{1+\rho^2}{\lambda^{n-1}} & -\frac{\rho}{\lambda^{\frac{2n-1}{2}}} \\ 0 & 0 & \dots & -\frac{\rho}{\lambda^{\frac{2n-1}{2}}} & \frac{1}{\lambda^n} \end{bmatrix} \quad (73)$$

where  $c > 0$ ,  $0 < \lambda \leq 1$ , and  $|\rho| < 1$ . Thus we have the following result on  $r_D^*$  and  $r_A^*$  for the DC kernel.

**Corollary 1** Consider the DC kernel (14). Then for  $\rho > 0$  with any  $N$ , or for  $\rho < 0$  with even  $N$ , we have  $r_D^* \neq r^\dagger$  and  $r_A^* \neq r^\dagger$ . In particular, for the TC kernel (15),  $r_D^* \neq r^\dagger$  and  $r_A^* \neq r^\dagger$  for any  $N$ .

### 5.2.2 More general nondiagonal kernel matrices

Now we consider more general nondiagonal kernel matrices and we start from a simple case when  $n = 2$ .

**Proposition 4** Consider the kernel matrix

$$P^{-1} = \begin{bmatrix} p_1 & -e \\ -e & p_2 \end{bmatrix} \quad (74)$$

where  $p_1 > 0, p_2 > 0$ , and  $e \neq 0$ . Then we have  $r_D^* \neq r^\dagger$  and  $r_A^* \neq r^\dagger$  for any  $N$ .

Then we consider nondiagonal kernel matrices with  $n \geq 2$  and  $n \in \mathbb{N}$ .

**Theorem 5** Suppose the kernel matrix  $P$  is positive definite. When  $N \geq 2n - 2$ , we have

$$r_D^* \neq r^\dagger$$

if at least one of the values  $\{\text{Tr}(Q(r^\dagger)^{-1}Q_i), 1 \leq i \leq n-1\}$  is nonzero and

$$r_A^* \neq r^\dagger$$

if at least one of the values  $\{\text{Tr}(Q(r^\dagger)^{-2}Q_i), 1 \leq i \leq n-1\}$  is nonzero.

**Corollary 2** Consider the DC kernel (14). Then we have  $r_D^* \neq r^\dagger$  and  $r_A^* \neq r^\dagger$ , if  $N \geq 2n - 2$ .

Theorems 4 to 5 and Corollaries 1 to 2 show that there exist cases such that  $r_D^*$  and  $r_A^*$  are no longer  $r^\dagger$ , implying that the information on the correlation of the impulse response indeed has influence on the input design.

Finally, one may wonder whether it is possible to claim that for all nondiagonal kernels,  $r_D^*$  and  $r_A^*$  are *not* equal to  $r^\dagger$ . Unfortunately, this claim is not true as can be seen from the following counter example.

**Example 1** Let  $\mathcal{E} = 1$ ,  $\sigma^2 = 1$ , and

$$P^{-1} = \begin{bmatrix} 1 & \frac{1}{2} & -\frac{1}{8} \\ \frac{1}{2} & 1 & -\frac{1}{2} \\ -\frac{1}{8} & -\frac{1}{2} & 1 \end{bmatrix}, \quad (75)$$

which has the eigenvalues 0.3526, 0.875, and 1.7724, and thus is positive definite, and yields

$$Q(r^\dagger)^{-1} = \begin{bmatrix} \frac{8}{15} & -\frac{2}{15} & 0 \\ -\frac{2}{15} & \frac{17}{30} & \frac{2}{15} \\ 0 & \frac{2}{15} & \frac{8}{15} \end{bmatrix}.$$

Then it can be shown that  $\text{Tr}(Q(r^\dagger)^{-1}Q_1) = 0$  and  $\text{Tr}(Q(r^\dagger)^{-1}Q_2) = 0$ , which implies that the gradient of  $\log \det(\sigma^2 Q(r)^{-1})$  at  $r = r^\dagger$  is zero and thus  $r_D^* = r^\dagger$  for nondiagonal kernel matrix (75) and for any  $N \geq 3$ .

## 6 Numerical Simulation

We illustrate by Monte Carlo simulations that the optimal input derived from the proposed method can improve the quality of the regularized FIR model estimate (10b) in contrast with the white noise input.

### 6.1 Test Systems

We first use the method in Chen et al. (2012); Pillonetto & Chiuso (2015) to generate 1000 30th order LTI systems. For each system, we truncate its impulse response at the order 50 and obtain a FIR model of order 50 accordingly, which is treated as the test system. In this way, we generate 1000 test systems.

### 6.2 Preliminary Data-Bank

For each test system, we generate a preliminary data record as follows, whose usage is to get a preliminary estimate of the kernel matrix  $P(\eta)$  and the noise variance  $\sigma^2$  necessary for the input design. We simulate each test system with a white Gaussian noise input with  $N = 50$  and  $\mathcal{E} = 10$ , treat the unknown inputs according to (29), and get the noise-free output with  $N = 50$ . The noise-free output is then corrupted by an additive white Gaussian noise. The signal-to-noise ratio (SNR), i.e., the ratio between the variance of the noise-free output and the noise, is uniformly distributed over  $[1, 10]$ . In this way, we get 1000 preliminary data records.

### 6.3 Optimal Input

For each test system and preliminary data record, we derive the optimal input as follows:

- 1) We consider FIR model with order  $n = 50$  and estimate the noise variance  $\sigma^2$  by LS method, as described in Chen et al. (2012); Goodwin et al. (1992) and the hyperparameter  $\eta$  of the TC kernel matrix  $P(\eta)$  given in (15) by the EB method (18).

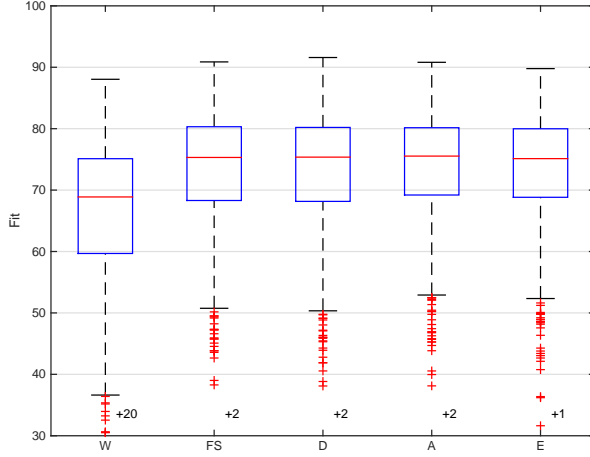


Fig. 1. Boxplots of the 1000 fits of the RLS estimates based on the preliminary data-bank and test databank.

- 2) Then for the obtained kernel matrix and noise variance, we derive the optimal input  $u_D^*$ ,  $u_A^*$ , and  $u_E^*$  by solving the input design problems (31), (32), (33), respectively, with  $N = 50$  and  $\mathcal{E} = 10$  and the proposed two-step procedure. For comparison, we also derive the optimal input in Fujimoto & Sugie (2016) with their gradient-based algorithm. It is worth to recall from (27) that the unknown inputs are set to zero in Fujimoto & Sugie (2016).

#### 6.4 Test Data-Bank

For each test system and each optimal input, we generate a test data record as follows, whose usage is to illustrate the efficacy of the input design. We simulate each test system with the optimal input, treat the unknown inputs according to (29), and get the noise-free output with  $N = 50$ . The noise-free output is then corrupted by an additive white Gaussian noise with the same variance as the additive white Gaussian noise in the preliminary data record. In this way, for each test system and each optimal input, we get 1000 test data records.

#### 6.5 Simulation Results

The simulation results are summarized in Table 1 and Fig. 1, where W, FS, D, A, E are used to denote the corresponding simulation results for different inputs, respectively, and W denotes the result corresponding to the white noise input in the preliminary data-bank. In particular, Table 1 shows the average model fits and Fig. 1 shows the boxplots of the 1000 model fits, where the model fit (Ljung, 2012) is defined as follows

$$\text{Fit} = 100 \times \left( 1 - \frac{\|\hat{\theta}_N^R - \theta_0\|}{\|\theta_0 - \tilde{\theta}_0\|} \right), \quad \tilde{\theta}_0 = \frac{1}{50} \sum_{k=1}^{50} g_k^0$$

where  $\hat{\theta}_N^R$  represents the regularized FIR model estimate (10b) corresponding to each data record, the regression matrix  $\Phi$  is constructed by the input according to (29), and the estimated noise variance and TC kernel matrix from the preliminary data record are used.

#### 6.6 Findings

In contrast with the white noise input, all optimal inputs including the one given in Fujimoto & Sugie (2016) and  $u_D^*$ ,  $u_A^*$ , and  $u_E^*$  given in this paper improve the average fit of the regularized FIR model estimate (10b) from 66 to 73 and are more robust. An intuitive explanation for this improvement is that the data records generated by the optimal inputs may have higher average SNRs than the preliminary data record generated by the white noise input. This guess is verified for the data records in this experiment. In addition, the fits of the estimates of the data records generated by all optimal inputs are quite close.

Table 1

Average fits of the regularized FIR model estimate (10b) given in Fig. 1 based on 1000 Monte Carlo runs.

W	FS	D	A	E
66.24	73.56	73.44	73.87	73.46

## 7 Conclusions

In this paper, the input design of kernel-based regularization method for LTI system identification has been investigated by minimizing the scalar measures of the Bayesian MSE matrix. Under suitable assumption on the unknown inputs and with the introduction of a quadratic map, the non-convex optimization problem associated with the input design problem is transformed into a convex optimization problem. Then based on the global minima of the convex optimization problem, the optimal input is derived by solving the inverse image of the quadratic map. For some special types of fixed kernels, some analytic results provide insights on the input design and also its dependence on the kernel structure.

## Appendix A

Appendix A contains the proof of the results in the paper, for which the technical lemmas are placed in Appendix B. The proofs of Propositions 1 and 2 are straightforward and thus omitted.

#### A.1 Proof of Theorem 1

First, we introduce the cyclic permutation matrix  $C \in \mathbb{R}^N$  defined by

$$C \triangleq \begin{bmatrix} 0 & I_{N-1} \\ 1 & 0 \end{bmatrix}, \quad (\text{A.1})$$

which has the following properties

$$C^i C^j = C^{i+j}, C^N = I_N, (C^i)^T = C^{N-i}. \quad (\text{A.2})$$

Note that  $L_0 = I_N$  and  $L_j = (C^j + (C^j)^T)/2$  is real symmetric for  $j = 1, \dots, n-1$ . Clearly,  $L_i$  commutes with  $L_j$  by using the properties (A.2) for  $i, j = 1, \dots, n-1$  and  $i \neq j$ . It follows from Theorem B1 in Appendix B that there exists an orthogonal matrix  $W$  such that

$$W^T L_j W = S_j, \quad j = 0, \dots, n-1, \quad (\text{A.3})$$

where  $S_j = \text{diag}[s_{j0}, \dots, s_{j,N-1}]$  is a real diagonal matrix. Therefore, we have

$$f_j(u) = u^T W S_j W^T u = \sum_{i=0}^{N-1} s_{ji} z_i^2 \quad (\text{A.4})$$

where  $z = [z_0, z_1, \dots, z_{N-1}]^T = W^T u$ . We obtain

$$f(u) = \underbrace{\begin{bmatrix} s_{00} & s_{01} & \cdots & s_{0,N-1} \\ s_{10} & s_{11} & \cdots & s_{1,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n-1,0} & s_{n-1,1} & \cdots & s_{n-1,N-1} \end{bmatrix}}_S \begin{bmatrix} z_0^2 \\ z_1^2 \\ \vdots \\ z_{N-1}^2 \end{bmatrix}. \quad (\text{A.5})$$

Since  $W^T W = I_N$ , the orthogonal linear mapping  $h_3(\cdot)$  has the property that the image is  $\mathcal{Z} = \{z | z^T z = \mathcal{E}\}$ . Then the image of  $h_2(\cdot)$  under  $\mathcal{Z}$  is the polytope  $\mathcal{X}$  given in (52) by the definition of  $h_2(\cdot)$ . Finally, the conclusion that  $\mathcal{F}$  is a polytope follows from that a linear transform of a polytope is still a polytope (Brøndsted, 1983, p. 18, 2.5). This completes the proof.

#### A.2 Proof of Theorem 2

We only prove the case for even  $N$ . The proof for odd  $N$  is similar and thus omitted.

For convenience, we place the construction of the matrices  $W$  and  $S$  in Lemma B1. So we first consider the claim  $\text{rank}(S) = \min(N/2 + 1, n)$ , which follows from that the vectors  $\xi_0, \xi_1, \dots, \xi_{\frac{N}{2}}$  are orthogonal to each other and  $\xi_j = \xi_{N-j}$  for  $j = 0, \dots, N$ .

Then note that the polyhedron  $\mathcal{X}$  in Theorem 1 can be written as a polytope

$$\mathcal{X} = \text{conv}\{d_0, d_1, \dots, d_{N-1}\}, \quad (\text{A.6})$$

where for  $j = 0, \dots, N-1$ ,  $d_j$  is the vector whose elements are zero except the  $j+1$ -element that is equal to  $\mathcal{E}$ . Since  $\cos(lj\varpi) = \cos((N-l)j\varpi)$  for any integer  $l$

and  $j = 0, \dots, N-1$ , the matrix  $S$  can be expressed in the form of (55).

Now we denote by  $\mathcal{C}$  the convex hull of the column vectors of  $S$  in (55) and we will show that  $\mathcal{F} = \mathcal{C}$ . On the one hand, note that

$$S d_j = S d_{N-j} = \mathcal{E} \xi_j(1:n), \quad j = 0, \dots, N/2. \quad (\text{A.7})$$

This implies that the image of (A.6) under the linear transform  $h_1(x) = Sx$  is included in  $\mathcal{C}$  and hence  $\mathcal{F} \subset \mathcal{C}$ . On the other hand, for each element  $r$  in  $\mathcal{C}$ , there exist  $a_j \geq 0, j = 0, \dots, N/2$  with  $\sum_{j=0}^{N/2} a_j = 1$  such that

$$r = \sum_{j=0}^{N/2} a_j \mathcal{E} \xi_j(1:n) = S \left( \sum_{j=0}^{N/2} a_j d_j \right) \in \mathcal{F}$$

where (A.7) is used. This concludes  $\mathcal{C} \subset \mathcal{F}$ . Therefore, we have  $\mathcal{F} = \mathcal{C}$ . This completes the proof.

#### A.3 Proof of Proposition 3

First, we show that  $r^\dagger \in \mathcal{F}$ . This can be done by finding a point  $x^0 \in \mathcal{X}$  such that  $Sx^0 = r^\dagger$ . We can choose  $x^0 = \mathcal{E}[1, \dots, 1]^T/N \in \mathcal{X}$ . Applying the property that  $\xi_0$  is orthogonal to the vectors  $\{\xi_j, 1 \leq j \leq N/2\}$  for even  $N$  or  $\{\xi_j, 1 \leq j \leq (N-1)/2\}$  for odd  $N$  yields  $Sx^0 = r^\dagger$ . Then we check the gradient of  $\log \det(\sigma^2 Q(r)^{-1})$  with respect to the last  $n-1$  variables of  $r$  is zero at  $r = r^\dagger$ . For ridge kernels, we have  $\sigma^2 P^{-1} = \sigma^2 I_n/c$  and accordingly  $Q(r^\dagger)$  is  $(\mathcal{E} + \sigma^2/c)I_n$ . Thus we find that

$$-\text{Tr}(Q(r^\dagger)^{-1} Q_i) = 0, \quad i = 1, \dots, n-1.$$

Since the problem (42) is strictly convex,  $r^\dagger$  is the unique stationary point and  $r_D^* = r^\dagger$ .

The proof for  $r_A^* = r^\dagger$  can be derived in a similar way.

#### A.4 Proof of Theorem 3

The proof of Theorem 3 is the same as that of Proposition 3 since  $Q(r^\dagger)$  is still diagonal in this case.

#### A.5 Proof of Theorem 4

First, we consider the case  $e_j > 0$  for  $j = 2, \dots, n$ . In this case, we have each element of  $Q(r^\dagger)^{-1} = (\mathcal{E} I_n + \sigma^2 P^{-1})^{-1}$  is positive by Theorem B3. This implies that each element of the gradient at  $r = r^\dagger$  is negative by (45) and the definition of  $Q_i$ . It follows that

$$[\nabla \log \det(\sigma^2 Q(r^\dagger)^{-1})]^T \xi_0(2:n) < 0,$$

since all elements of  $\xi_0(2 : n)$  are equal to one. This violates Lemma B2 in Appendix B, which gives the necessary and sufficient condition for  $r_D^* = r^\dagger$ . Thus  $r_D^* \neq r^\dagger$  for this case.

Then we consider the case where  $e_j < 0$  for  $j = 2, \dots, n$  and  $N$  is even. In this case, we have  $(Q(r^\dagger)^{-1})_{ij} > 0$  when  $|i - j|$  is even and  $(Q(r^\dagger)^{-1})_{ij} < 0$  when  $|i - j|$  is odd by Theorem B3. This yields that the sign of the  $i$ -th element  $\frac{\partial \log \det(\sigma^2 Q(r^\dagger)^{-1})}{\partial r_i}$  of the gradient  $\nabla \log \det(\sigma^2 Q(r^\dagger)^{-1})$  is positive for odd  $i$  and negative for even  $i$ . This shows that

$$[\nabla \log \det(\sigma^2 Q(r^\dagger)^{-1})]^T \xi_{N/2}(2 : n) < 0$$

since the  $i$ -th element of  $\xi_{N/2}(2 : n)$  is  $(-1)^i$ . This violates the necessary and sufficient condition for  $r_D^* = r^\dagger$ , and thus  $r_D^* \neq r^\dagger$  for this case.

The proof of  $r_A^* \neq r^\dagger$  is similar and thus omitted.

#### A.6 Proof of Corollary 1

The proof is trivial by noting (73) and Theorem 4.

#### A.7 Proof of Proposition 4

It follows that

$$Q(r^\dagger)^{-1} = \frac{1}{\det(Q(r^\dagger))} \begin{bmatrix} \mathcal{E} + \sigma^2 p_2 & \sigma^2 e \\ \sigma^2 e & \mathcal{E} + \sigma^2 p_1 \end{bmatrix}.$$

Then we have  $-\text{Tr}(Q(r^\dagger)^{-1} Q_1) = -2\sigma^2 e / \det(Q(r^\dagger))$ . In this case, we have  $\xi_j(2 : n) = \xi_j(2 : 2) = \cos(j\varpi)$ . As a result, there always exists an index  $j$  such that  $e$  and  $\cos(j\varpi)$  have the same sign for  $j = 0, \dots, N/2$  when  $N$  is even and for  $j = 0, \dots, (N-1)/2$  when  $N$  is odd. One obtains that  $-e \cos(j\varpi) < 0$  for this  $j$ , which violates (B.14). Therefore,  $r_D^* \neq r^\dagger$ . The proof of  $r_A^* \neq r^\dagger$  is similar and is omitted.

#### A.8 Proof of Theorem 5

It follows from Lemma B3 in Appendix B that 0 is an interior point of  $\mathcal{F}' = \{r' \in \mathbb{R}^{n-1} | [\mathcal{E}, r'^T]^T \in \mathcal{F}\}$  when  $N \geq 2n - 2$ . Since  $P^{-1}$  is nondiagonal, we have  $Q(r^\dagger)^{-1}$  is also nondiagonal. The condition that at least one of the values  $\text{Tr}(Q(r^\dagger)^{-1} Q_i)$ ,  $i = 1, \dots, n - 1$ , is nonzero implies that  $r^\dagger$  is not a stationary point of the problem (42). This means that  $r_D^* \neq r^\dagger$ . The proof for  $r_A^* \neq r^\dagger$  is similar and thus omitted.

#### A.9 Proof of Corollary 2

For the DC kernel (14), by Theorem B3 as used in the proof of Theorem 4 we have

- 1) When  $\rho > 0$ ,  $(Q(r^\dagger)^{-1})_{ij} > 0$  and  $(Q(r^\dagger)^{-2})_{ij} > 0$ .
- 2) When  $\rho < 0$ ,  $(Q(r^\dagger)^{-1})_{ij} > 0$  and  $(Q(r^\dagger)^{-2})_{ij} > 0$  if  $|i - j|$  is even and  $(Q(r^\dagger)^{-1})_{ij} < 0$  and  $(Q(r^\dagger)^{-2})_{ij} < 0$  if  $|i - j|$  is odd.

This implies that  $\text{Tr}(Q(r^\dagger)^{-1} Q_i) \neq 0$  and  $\text{Tr}(Q(r^\dagger)^{-2} Q_i) \neq 0$  for all indices  $i = 1, \dots, n - 1$ . As a result, the proposition is true by Theorem 5.

### Appendix B

This appendix contains the technical lemmas used in the proof in Appendix A.

**Theorem B1** (Jiang & Li, 2016, Theorem 9) *Real symmetric matrices  $\{I_N, A_1, \dots, A_m\}$  with  $A_i \in \mathbb{R}^{N \times N}$  and  $i = 1, \dots, m$  are simultaneously diagonalizable via an orthogonal congruent matrix if and only if  $A_i$  commutes with  $A_j$  for  $i, j = 1, 2, \dots, m$  and  $i \neq j$ .*

**Theorem B2** (Gray (2006, Theorem 3.1); Tee (2007)) *Denote the circulant matrix  $B$  generated by a row vector  $b = [b_0, b_1, \dots, b_{N-1}]$  by*

$$B = \text{circ}(b) \triangleq \begin{bmatrix} b_0 & b_1 & \cdots & b_{N-2} & b_{N-1} \\ b_{N-1} & b_0 & \cdots & b_{N-3} & b_{N-2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ b_2 & b_3 & \cdots & b_0 & b_1 \\ b_1 & b_2 & \cdots & b_{N-1} & b_0 \end{bmatrix}. \quad (\text{B.8})$$

*Then  $B$  has unit eigenvectors*

$$v^{(m)} = \frac{1}{\sqrt{N}} [1, \exp(-\iota \varpi m), \dots, \exp(-\iota \varpi (N-1)m)]^T, \\ m = 0, \dots, N-1,$$

*where  $\varpi = 2\pi/N$  and  $\iota$  is the imaginary unit ( $\iota^2 = -1$ ), and the corresponding eigenvalues*

$$\tau^{(m)} = \sum_{k=0}^{N-1} b_k \exp(-\iota m k \varpi) \\ = b[1, \exp(-\iota m \varpi), \dots, \exp(-\iota m (N-1) \varpi)]^T$$

*and can be expressed by*

$$B = A \text{diag}([\tau^{(0)}, \dots, \tau^{(N-1)}]) A^H$$



where  $A = [v^{(0)}, \dots, v^{(N-1)}]$  is unitary and  $A^H$  denotes the complex conjugate transpose of  $A$ .

**Theorem B3** Let  $A$  be a symmetric and positive definite tridiagonal matrix of dimension  $n$ ,

$$A = \begin{bmatrix} a_1 & -b_2 & & & \\ -b_2 & a_2 & -b_3 & & \\ & \ddots & \ddots & \ddots & \\ & & -b_{n-1} & a_{n-1} & -b_n \\ & & & -b_n & a_n \end{bmatrix} \quad (\text{B.9})$$

and denote the  $(i, j)$ -element of  $A^{-1}$  by  $(A^{-1})_{ij}$ . Then we have  $(A^{-1})_{ij} > 0$  when  $b_l > 0$  for  $l = 2, \dots, n$ , while  $(A^{-1})_{ij} < 0$  if  $|i - j|$  is odd and  $(A^{-1})_{ij} > 0$  if  $|i - j|$  is even when  $b_l < 0$  for  $l = 2, \dots, n$ .

**Proof.** The result is obtained by applying Theorem 2.3 of Meurant (1992) when  $A$  is positive definite.  $\blacksquare$

**Theorem B4** (Boyd & Vandenberghe, 2004, Section 4.2.3, page 139) Let  $\varphi : \mathbb{R}^m \rightarrow \mathbb{R}$  be a differentiable convex function, and let  $\mathcal{B} \subset \mathbb{R}^m$  be a nonempty closed convex set. Consider the problem

$$\text{minimize } \varphi(h) \quad \text{subject to } h \in \mathcal{B}. \quad (\text{B.10})$$

A vector  $h^*$  is optimal for this problem if and only if  $h^* \in \mathcal{B}$  and

$$\nabla \varphi(h^*)^T (y - h^*) \geq 0 \quad \text{for all } y \in \mathcal{B} \quad (\text{B.11})$$

where

$$\nabla \varphi(\cdot) \triangleq \left[ \frac{\partial \varphi(\cdot)}{\partial h_1}, \dots, \frac{\partial \varphi(\cdot)}{\partial h_m} \right]^T$$

is the gradient of  $\varphi(\cdot)$  with respect to  $h = [h_1, \dots, h_m]^T$ .

**Lemma B1** The matrices  $\{L_j, j = 0, \dots, n-1\}$  can be diagonalized simultaneously by the matrix  $W$ , i.e.,

$$W^T L_j W = S_j = \text{diag}(\xi_j), \quad j = 0, \dots, n-1. \quad (\text{B.12})$$

**Proof.** First, we consider the case where  $N$  is even. It follows that  $L_0 = I_N = \text{circ}([1, 0, \dots, 0])$  and  $L_j = \text{circ}(l_j/2)$ ,  $j = 1, \dots, n-1$ , where  $l_j$  is the  $N$ -dimensional row vector whose elements are zero except that the  $j+1$ -th and  $N-j+1$ -th elements are one. Then by Theorem B2, the eigenvalues of  $L_j$  are

$$\begin{aligned} \tau^{(m,j)} &= l_j [1, \exp(-im\varpi), \dots, \exp(-im(N-1)\varpi)]^T / 2 \\ &= \exp(-imj\varpi) / 2 + \exp(-im(N-j)\varpi) / 2 \\ &= \cos(mj\varpi), \quad m=0, \dots, N-1, \end{aligned}$$

the eigenvectors  $\{v^{(m)}, m = 0, \dots, N-1\}$  is an orthonormal basis for all matrices  $L_j, j = 0, \dots, n-1$ . Moreover, we have  $v^{(0)} = \xi_0 / \sqrt{N}$  and  $v^{(N/2)} = \xi_{N/2} / \sqrt{N}$  are real, and  $\tau^{(m,j)} = \tau^{(N-m,j)}$ , which implies that the eigenvectors  $v^{(m)}$  and  $v^{(N-m)}$  of  $L_j$  correspond to the same eigenvalue  $\cos(mj\varpi)$ . Thus the linear combinations  $(v^{(m)} + v^{(N-m)})/2 = \xi_m / \sqrt{N}$  and  $(v^{(N-m)} - v^{(m)})/2 = \zeta_m / \sqrt{N}$  of the eigenvectors  $v^{(m)}$  and  $v^{(N-m)}$  corresponding to the  $m$ -th and  $(N-m)$ -th eigenvalue  $\cos(mj\varpi)$  of  $L_j$  are real and orthogonal. Further we have  $|\xi_m| = |\zeta_m| = \sqrt{N/2}$ , and thus

$$\left\{ \xi_0, \sqrt{2}\xi_1, \dots, \sqrt{2}\xi_{N/2-1}, \xi_{N/2}, \sqrt{2}\zeta_{N/2-1}, \dots, \sqrt{2}\zeta_1 \right\} / \sqrt{N}$$

is a group of real orthonormal eigenvectors of  $L_j$  and the corresponding eigenvalues are  $\{1, \cos(j\varpi), \dots, \cos((N-1)j\varpi)\}$ , which are actually the elements of  $\xi_j$ . This completes the proof for the case for even  $N$ . The proof for odd  $N$  is similar and thus omitted.  $\blacksquare$

**Lemma B2** The vector  $r^\dagger \in \mathcal{F}$  is the solution of the problem

$$\text{minimize } \varphi(r) \quad \text{subject to } r \in \mathcal{F} \quad (\text{B.13})$$

where  $\varphi(r)$  represents  $\log \det(\sigma^2 Q(r)^{-1})$  or  $\text{Tr}(\sigma^2 Q(r)^{-1})$ , if and only if

$$[\nabla \varphi(r^\dagger)]^T \xi_j(2:n) \geq 0 \quad (\text{B.14})$$

with  $j = 0, \dots, N/2$  for even  $N$  or  $j = 0, \dots, (N-1)/2$  for odd  $N$ , where  $\nabla \varphi(r^\dagger) = \left[ \frac{\partial \varphi(r^\dagger)}{\partial r_1}, \dots, \frac{\partial \varphi(r^\dagger)}{\partial r_{n-1}} \right]^T$ .

**Proof.** Note that the first element of each member in  $\mathcal{F}$  is  $\mathcal{E}$  and hence the first element of the difference  $r - r^\dagger$  is zero for any  $r \in \mathcal{F}$ . In addition, we have the last  $n-1$  elements of  $r^\dagger$  are zero. By applying Theorem B4, we see that  $r^\dagger$  is the solution of (B.13) if and only if

$$[\nabla \varphi(r^\dagger)]^T (\tilde{r} - 0) \geq 0, \quad (\text{B.15})$$

where  $\tilde{r}$  is any convex linear combination of the points  $\{\xi_j(2:n), j = 0, \dots, N/2\}$  when  $N$  is even, namely,  $\tilde{r} = \sum_{j=0}^{N/2} a_j \xi_j(2:n)$  with  $a_j \geq 0$  and  $\sum_{j=0}^{N/2} a_j = 1$ . Clearly, the condition (B.15) is equivalent to (B.14) for  $j = 0, \dots, N/2$ . The case when  $N$  is odd can be proved in a similar way and thus omitted.  $\blacksquare$

**Lemma B3** The zero vector  $0 \in \mathbb{R}^{n-1}$  is an interior point of  $\mathcal{F}' = \{r' \in \mathbb{R}^{n-1} | [\mathcal{E}, r'^T]^T \in \mathcal{F}\}$  if  $N \geq 2n-2$ .

**Proof.** This conclusion is proved by considering two cases, respectively: 1) for even  $N$  with  $N \geq 2n-2$ ; 2) for odd  $N$  with  $N \geq 2n-1$ . Let us explore the first case. It is clear that the dimension of  $\mathcal{F}'$  is equal to or less than

$n - 1$ . It follows from Theorem 2 that  $\text{rank}(S) = n$ . This means that the dimension of  $\mathcal{F}'$  is  $n - 1$  and  $\mathcal{F}'$  consists of all convex combinations of affinely independent  $n$  vectors from  $\{\mathcal{E}\xi_j(2:n), j = 0, \dots, N/2\}$  (Brøndsted, 1983, p. 14, Corollary 2.4). Now we will prove the result by contradiction. Assume that 0 is not an interior point of  $\mathcal{F}'$ . Thus 0 must be located on a facet of  $\mathcal{F}'$  and the dimension of this facet is  $n - 2$  since the dimension of  $\mathcal{F}'$  is  $n - 1$  and hence 0 is a convex combination of at most  $n - 1$  affinely independent vectors from  $\{\mathcal{E}\xi_j(2:n), j = 0, \dots, N/2\}$  (Brøndsted, 1983, Corollary 2.4). In contrast, 0 can be expressed by a convex combination in the following way

$$0 = \frac{1}{N}(\mathcal{E}S(2:n, :))\xi_0 = \frac{1}{N}(\mathcal{E}\xi_0(2:n) + \mathcal{E}\xi_{\frac{N}{2}}(2:n)) + \frac{2}{N} \sum_{j=1}^{\frac{N}{2}-1} (\mathcal{E}\xi_j(2:n)) \quad (\text{B.16})$$

due to the orthogonal vectors  $\{\xi_j, j = 0, 1, \dots\}$  and the expression (55) of  $S$ . Further, the expression (B.16) can be rewritten as a convex combination of affinely independent  $n$  vectors from  $\{\xi_j(2:n), j = 0, \dots, N/2\}$  and all the weights are positive since the remaining  $N/2 + 1 - n$  vectors of  $\{\xi_j(2:n), j = 0, \dots, N/2\}$  are convex combinations of these  $n$  affinely independent vectors. This contradiction means that the proposition is true. The proof for odd  $N$  with  $N \geq 2n - 1$  is similar and thus omitted. ■

## References

- Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.
- Brøndsted, A. (1983). *An introduction to convex polytopes*. New York, Inc: Springer-Verlag.
- Carli, F. P., Chen, T., & Ljung, L. (2017). Maximum entropy kernels for system identification. *IEEE Transactions on Automatic Control*, 62, 1471–1477.
- Chen, T., Andersen, M. S., Ljung, L., Chiuso, A., & Pillonetto, G. (2014). System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques. *IEEE Transactions on Automatic Control*, 59, 2933–2945.
- Chen, T., Ardeschiri, T., Carli, F. P., Chiuso, A., Ljung, L., & Pillonetto, G. (2016). Maximum entropy properties of discrete-time first-order stable spline kernel. *Automatica*, 66, 34–38.
- Chen, T., & Ljung, L. (2017). On kernel design for regularized lti system identification. *arXiv preprint arXiv:1612.03542*, .
- Chen, T., Ohlsson, H., & Ljung, L. (2012). On the estimation of transfer functions, regularizations and gaussian processes—revisited. *Automatica*, 48, 1525–1535.
- Chiuso, A. (2016). Regularization and bayesian learning in dynamical systems: Past, present and future. *Annual Reviews in Control*, 41, 24–38.
- Fujimoto, Y., & Sugie, T. (2016). Informative input design for kernel-based system identification. In *Proceedings of IEEE Conference on Decision and Control* (pp. 4636–4639). IEEE.
- Gevers, M. (2005). Identification for control: From the early achievements to the revival of experiment design. *European journal of control*, 11, 335–352.
- Goodwin, G. C., Gevers, M., & Ninness, B. (1992). Quantifying the error in estimated transfer functions with application to model order selection. *IEEE Transactions on Automatic Control*, 37, 913–928.
- Goodwin, G. C., & Payne, R. L. (1977). *Dynamic system identification: experiment design and data analysis*. New York: Academic press.
- Grant, M. C., & Boyd, S. P. (2016). Cvx: Matlab software for disciplined convex programming, version 2.1. Available from <http://cvxr.com/cvx/>.
- Gray, R. M. (2006). *Toeplitz and circulant matrices: A review*. Now Publishers, Inc.
- Hildebrand, R., & Gevers, M. (2003). Identification for control: optimal input design with respect to a worst-case  $\nu$ -gap cost function. *SIAM Journal on Control and Optimization*, 41, 1586–1608.
- Hjalmarsson, H. (2005). From experiment design to closed-loop control. *Automatica*, 41, 393–438.
- Hjalmarsson, H. (2009). System identification of complex and structured systems. *European journal of control*, 15, 275–310.
- Jansson, H., & Hjalmarsson, H. (2005). Input design via lmis admitting frequency-wise model specifications in confidence regions. *IEEE transactions on Automatic Control*, 50, 1534–1549.
- Jiang, R., & Li, D. (2016). Simultaneous diagonalization of matrices and its applications in quadratically constrained quadratic programming. *SIAM Journal on Optimization*, 26, 1649–1668.
- Ljung, L. (1999). *System identification: Theory for the user*. Upper Saddle River, NJ: Prentice-Hall.
- Ljung, L. (2012). *System Identification Toolbox for Use with MATLAB*. (8th ed.). Natick, MA: The MathWorks, Inc.
- Ljung, L., Singh, R., & Chen, T. (2015). Regularization features in the system identification toolbox. In *Proceedings of the IFAC Symposium on System Identification* (pp. 745–750). Beijing, China.
- Marconato, A., Schoukens, M., & Schoukens, J. (2016). Filter-based regularisation for impulse response modelling. *IET Control Theory & Applications*, 11, 194–204.
- Mehra, R. (1974). Optimal input signals for parameter estimation in dynamic systems—survey and new results. *IEEE Transactions on Automatic Control*, 19, 753–768.
- Meurant, G. (1992). A review on the inverse of symmetric tridiagonal and block tridiagonal matrices. *SIAM Journal on Matrix Analysis and Applications*, 13, 707–728.
- Mu, B., Chen, T., & Ljung, L. (2017). On asymptotic properties of hyperparameter estimators for

- kernel-based regularization methods. *arXiv preprint arXiv:1707.00407*, .
- Pillonetto, G., Chen, T., Chiuso, A., De Nicolao, G., & Ljung, L. (2016). Regularized linear system identification using atomic, nuclear and kernel-based norms: The role of the stability constraint. *Automatica*, *69*, 137–149.
- Pillonetto, G., & Chiuso, A. (2015). Tuning complexity in regularized kernel-based regression and linear system identification: The robustness of the marginal likelihood estimator. *Automatica*, *58*, 106–117.
- Pillonetto, G., Chiuso, A., & De Nicolao, G. (2011). Prediction error identification of linear systems: a non-parametric gaussian regression approach. *Automatica*, *47*, 291–305.
- Pillonetto, G., & De Nicolao, G. (2010). A new kernel-based approach for linear system identification. *Automatica*, *46*, 81–93.
- Pillonetto, G., Dinuzzo, F., Chen, T., De Nicolao, G., & Ljung, L. (2014). Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, *50*, 657–682.
- Söderström, T., & Stoica, P. (1989). *System identification*. Prentice Hall International.
- Tee, G. J. (2007). Eigenvectors of block circulant and alternating circulant matrices. *New Zealand Journal of Mathematics*, *36*, 195–211.
- Zarrop, M. B. (1979). *Optimal experiment design for dynamic system identification*. Berlin Heidelberg: Springer-Verlag.
- Zorzi, M., & Chiuso, A. (2017). The harmonic analysis of kernel functions. *arXiv preprint arXiv:1703.05216*, .