Information Inequality for Estimation of Transfer Functions: main results [★],[★★]

Tzvetan Ivanov [∗] **Brian D.O. Anderson** [∗∗] **P.-A. Absil** [∗]
Michel Gevers [∗]

[∗] *Center for Systems Engineering and Applied Mechanics (CESAME)*
Universite Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium.
(tzvetan.ivanov@uclouvain.be, michel.gevers@uclouvain.be,
absil@inma.ucl.ac.be)
[∗∗] *National ICT Australia and the Australian National University*
Research School of Information Sciences & Engineering, Canberra,
ACT 2601, Australia (brian.anderson@anu.edu.au)

Abstract: In this paper we derive a canonical lower bound for the autocovariance function of any unbiased transfer-function estimator. As a generalization of the Cramér-Rao bound, the Cramér-Rao kernel that we define can be derived without parametrizing the model set. The Cramér-Rao kernel is thus one of the cornerstones for experiment design formulations that do not depend on the choice of coordinates.

Keywords: System Identification, Metric, Confidence Region, Autocovariance, Reproducing Kernel, Fisher Information, Information Geometry

1. INTRODUCTION

Parameter estimation enjoys a rich set of features which makes it attractive in the context of system identification (SYSID). One of the main features is that parameter estimation, when used in a probabilistic framework, allows users to quantify their confidence in how closely the estimated parameter matches the true one. For example in experiment design one uses classical results, such as the Cramér-Rao lower bound (CRLB), in order to tune the experimental conditions in such a way that the achieved confidence region matches some performance specification. However, finding the optimal experimental conditions can be challenging if the performance specification is not directly given in the parameter space. For example, if the goal is to estimate a transfer function, then the performance specification is typically given in terms a of maximal tolerable distance between the true transfer function and the estimated transfer function and not in terms of a maximal tolerable distance between the corresponding parameter vectors. In order to execute the experiment design one can often use the fact that they are related via a bijective function – the parametrization.

Additionally, the notion of distance between the true transfer function and the estimated transfer function de-

pends on the application, and is often induced by norms on the embedding space such as the L_∞-norm or L_2-norm. In general, and perhaps regrettably, natural distance functions on the model set lead to complicated, non-linear, distance functions on the parameter space.

We are therefore interested to circumvent the usage of parametrizations and replace the main players in the optimal experiment design problem. To be concrete let $C(\hat{\theta}) \in \mathbf{R}^{d \times d}$ denote the covariance matrix of an unbiased estimator $\hat{\theta}$ of a parameter $\theta^* \in \Theta \subseteq \mathbf{R}^d$, and assume the Fisher-information matrix, call it $J(\theta^*) \in \mathbf{R}^{d \times d}$, is a nonsingular matrix. Then the problem is how to replace the CRLB, i.e., the matrix inequality

$$C(\hat{\theta}) \geq J(\theta^*)^{-1}, \qquad (1)$$

by an inequality between two objects which are defined purely in terms of the model set and the probabilistic assumption, i.e., in a manner which does not involve parametrizations at all. In most cases then, we are seeking a replacement of (1) by an inequality somehow reflecting transfer functions.

One part of this problem has been solved in Rao (1945), where one showed that the information matrix $J(\theta)$ at a fixed parameter $\theta^* \in \Theta$ satisfies

$$[J(\theta^*)]_{ij} = \left(\frac{\partial \Pi(\theta^*)}{\partial \theta_j}, \frac{\partial \Pi(\theta^*)}{\partial \theta_i} \right)_{P_*}, \qquad (2)$$

where $(\cdot, \cdot)_{P_*}$ denotes an inner-product at the tangent space of the model set $\mathcal{P}$ at the unique model $P_* \in \mathcal{P}$ such that $P_* = \Pi(\theta^*)$ holds for the parametrization $\Pi : \Theta \to \mathcal{P}$. The key point is that this inner-product is actually independent of the chosen parametrization and, when viewed as a function of P_*, defines a metric, the so called *information metric*, on the model set, cf. Amari

and Nagaoka (2001). This observation alone, however, does not suffice to solve the original problem which requires parametrization independent definitions for the covariance matrix of the estimated quantity. The inversion of the information matrix and also the standard matrix ordering used in (1) has to be redefined in an appropriate way.

The contribution of this paper is the derivation of a coordinate-free version of the CRLB (1) for the case where:

- The elements $P \in \mathcal{P}$ of the model set $\mathcal{P}$ are complex-valued functions $P : \mathbb{T} \to \mathbf{C}$, defined on the unit circle $\mathbb{T}$ of the complex plane $\mathbf{C}$, i.e., discrete-time transfer functions.
- The model set $\mathcal{P}$ is a d-dimensional differentiable manifold over the real numbers $\mathbf{R}$, i.e., the set of transfer functions admits a smooth parametrization defined on a subset of $\mathbf{R}^d$.
- The information metric $(\cdot, \cdot)_P$ is non-singular, i.e., the data set is sufficiently rich that unique identification can be achieved.

We show that for every function estimator $\hat{P}$ such that

$$\mathbb{E}_{P_*}[\hat{P}(z)] = P_*(z) \quad \text{for all} \quad z \in \mathbb{T}, P_* \in \mathcal{P}, \quad (3)$$

(i.e., $\hat{P}$ is unbiased), *the corresponding autocovariance function admits a canonical lower bound designated K_{P_*} below.*

The autocovariance function is the two variable function defined for all $z, w \in \mathbb{T}$ via

$$\text{Cov}_{P_*}(\hat{P})(z, w) := \mathbb{E}[\overline{(\hat{P}(z) - P_*(z))} \cdot (\hat{P}(w) - P_*(w))]. \quad (4)$$

By lower bound we mean an inequality of the form

$$\sum_{k,l=1}^M \eta_k \overline{\eta_l} \cdot \left(\text{Cov}_{P_*} \hat{P}(z_k, z_l) - K_{P_*}(z_k, z_l) \right) \geq 0, \quad (5)$$

which holds for every integer $M \geq 0$, every set of evaluation points $z_1, \ldots, z_M \in \mathbb{T}$, and every vector $\eta \in \mathbf{C}^M$. The inequality (5) simply means

$$\text{Cov}_{P_*}(\hat{P}) \geq_{\text{EH}} K_{P_*}, \quad (6)$$

where $\geq_{\text{EH}}$ denotes the E.H. Moore ordering Moore (1916).

We show that the classical CRLB corresponds to the case where K_{P_*} is the *reproducing kernel* (cf. Appendix A) of the tangent-space of the model set at P_* with respect to the information metric $(\cdot, \cdot)_{P_*}$ defined in (2). Due to its fundamental importance in variance quantification we call K_{P_*} the Cramér-Rao kernel.

The paper is structured as follows. In Section 2 and 3 we develop the Cramér-Rao kernel lower bound starting from a parametric setup. In Section 4 we prove reparametrization invariance of the previous results which will allow us to replace them, in Section 5, by purely geometric definitions which do not involve parametrizations. In Section 6 we demonstrate the effectiveness of our approach by revisiting asymptotic sample size analysis in system identification for the special model set $\mathcal{P} = \text{Rat}(n)$ which consists of all transfer functions with McMillan degree n. In Section 7 we give a sample application arising from performance specifications used in robust control. In Section 8 we conclude.

2. THE CRAMÉR-RAO KERNEL

In this section we shall derive the Cramér-Rao kernel which plays the role of the inverse information matrix in the context of function estimation. To highlight the connections with the classical parametric approach in Lemma 2.1 we derive the Cramér-Rao kernel given a fixed parametrization Π of the function space $\mathcal{P}$.

Before we continue, we first introduce the setup and fix the notation for the rest of the paper:

- parameter-space $\Theta \subseteq \mathbf{R}^d$, parameter-vector $\theta \in \Theta$,
- sample-space $X \subseteq \mathbf{R}^q$, random sample [1] $\text{x} \in X$,
- parametrized set of probability density functions
$$\mathcal{D} = \{p_\theta : X \to \mathbf{R}_{\geq 0} \,|\, \theta \in \Theta\}. \quad (7)$$
- expectation $\mathbb{E}_p[f(\text{x})]$ of a particular measurement $f : X \to \mathbf{C}$ w.r.t. to a random sample $\text{x} \sim p$ with density $p \in \mathcal{D}$,
- parametrized model set of transfer functions
$$\mathcal{P} = \{P : \mathbb{T} \to \mathbf{C} \,|\, P = \Pi(\theta), \theta \in \Theta\}, \quad (8)$$
- $z \in \mathbb{T}$ evaluation point, evaluated parametrization
$$\Pi_z : \Theta \to \mathbf{C} \quad \text{with} \quad \Pi_z(\theta) := (\Pi(\theta))(z). \quad (9)$$
- parameter estimator $\hat{\theta}(\text{x}) \in \Theta$, function estimator
$$\hat{P}_{\text{x}} \in \mathcal{P} \quad \text{with} \quad \hat{P}_{\text{x}}(z) = \Pi_z(\hat{\theta}(\text{x})). \quad (10)$$
- unbiasedness for function estimators
$$\mathbb{E}_{p_\theta}[\hat{P}_{\text{x}}(z)] = \Pi_z(\theta) \quad \text{for all} \quad z \in \mathbb{T}, \theta \in \Theta. \quad (11)$$
- autocovariance of an unbiased estimator $\hat{P}$
$$\text{Cov}_{p_\theta}(\hat{P})(z, w) :=$$
$$\mathbb{E}_{p_\theta}[\overline{(\hat{P}_{\text{x}}(z) - \Pi_z(\theta))} \cdot (\hat{P}_{\text{x}}(w) - \Pi_w(\theta))], \quad (12)$$
for all $z, w \in \mathbb{T}$ and $\theta \in \Theta$.

We shall make the following regularity assumptions:

R1) The model set $\mathcal{P}$ is a differentiable manifold, the parameter space Θ is open, and Π is a smooth parametrization of $\mathcal{P}$.

R2) For every fixed $x \in X$ the function $\Theta \to \mathbf{R}_{\geq 0}, \theta \mapsto p_\theta(x)$ is smooth. Similarly for every fixed $z \in \mathbb{T}$ the function $\Theta \to \mathbf{C}, \theta \mapsto \Pi_z(\theta)$ is well-defined and smooth. [2]

R3) For every $\theta^* \in \Theta$ the information matrix J_{θ^*} given by
$$[J_{\theta^*}]_{ij} = \mathbb{E}_{p_{\theta^*}} \left[\frac{\partial \log p_{\theta^*}(\text{x})}{\partial \theta_j} \cdot \frac{\partial \log p_{\theta^*}(\text{x})}{\partial \theta_i} \right], \quad (13)$$
$i, j = 1, \ldots d$, is non-singular.

R1) states that the parameter vector θ is globally-identifiable, i.e., that $\Pi(\theta) = \Pi(\theta^*)$ implies that $\theta = \theta^*$ Gevers et al. (2009). R3) ensures that the samples are sufficiently rich to distinguish between parameters.

---

[1] A random sample is like a sequence of measurements on the system, for example a sequence of input and output values u_k, y_k, $k = 1, 2, ..., N$ for a linear plant with outputs contaminated by additive noise, defined by an equation like $y = Pu + v$ the symbols having obvious meaning.

[2] In particular we require that for all $\theta \in \Theta$ the function $P = \Pi(\theta)$ is defined for all $z \in \mathbb{T}$. If P is a rational function this is equivalent to saying that P has no pole on the unit circle.

Remark 2.1. Due to assumption R1), we may abuse notation and write $\mathbb{E}_P$ and Cov_P instead of $\mathbb{E}_p$ and Cov_p whenever there holds $p = p_\theta$ and $P = \Pi(\theta)$.

Lemma 2.1. Let $\hat{P}_{\mathrm{x}} \in \mathcal{P}$ denote an unbiased estimator in the sense of (11); then (5) holds with

$$K_{P_*}(z,w) = \sum_{i,j=1}^{d} \left[J_{\theta^*}^{-1} \right]_{ji} \cdot \frac{\partial \Pi_z(\theta^*)}{\partial \theta_i} \, \overline{\frac{\partial \Pi_w(\theta^*)}{\partial \theta_j}}, \quad (14)$$

where $P_* = \Pi(\theta^*)$. We call K_{P_*} the J_{θ^*}-*induced Cramér-Rao kernel* at the model P_*.

Proof. We first note that

$$\sum_{k,l=1}^{M} \eta_k \overline{\eta_l} \cdot \mathrm{Cov}_{P_*} \hat{P}(z_k, z_l) = \mathbb{E}_{P_*} |\hat{\alpha}(\mathrm{x}) - \alpha(\theta^*)|^2, \quad (15)$$

holds with

$$\hat{\alpha}(\mathrm{x}) = \sum_{k=1}^{M} \bar{\eta}_k \hat{P}_{\mathrm{x}}(z_k),$$
$$\alpha(\theta) = \sum_{k=1}^{M} \bar{\eta}_k \Pi_{z_k}(\theta).$$

Since $\hat{P}_{\mathrm{x}}$ is unbiased we have $\mathbb{E}_{P_*}[\hat{\alpha}(\mathrm{x})] = \alpha(\theta^*)$. By the classical CRLB there holds

$$\mathbb{E}_{P_*} |\hat{\alpha}(\mathrm{x}) - \alpha(\theta^*)|^2 \geq \sum_{i,j=1}^{d} [J_{\theta^*}^{-1}]_{ji} \frac{\partial \alpha(\theta^*)}{\partial \theta_i} \overline{\frac{\partial \alpha(\theta^*)}{\partial \theta_j}}$$

$$= \sum_{i,j=1}^{d} [J_{\theta^*}^{-1}]_{ji} \frac{\partial}{\partial \theta_i} \sum_{k=1}^{M} \bar{\eta}_k \Pi_{z_k}(\theta^*) \cdot \overline{\frac{\partial}{\partial \theta_j} \sum_{l=1}^{M} \bar{\eta}_l \Pi_{z_l}(\theta^*)}$$

$$= \sum_{k,l=1}^{M} \eta_k \bar{\eta}_l \sum_{i,j=1}^{d} \left[J_{\theta^*}^{-1} \right]_{ji} \cdot \frac{\partial \Pi_{z_k}(\theta^*)}{\partial \theta_i} \, \overline{\frac{\partial \Pi_{z_l}(\theta^*)}{\partial \theta_j}}$$

$$\sum_{k,l=1}^{M} \eta_k \bar{\eta}_l K_{P_*}(z_k, z_l),$$

which, together with equation (15), is equivalent to (5). $\square$

3. THE REPRODUCING PROPERTY

In the previous section we saw that the J_{θ^*}-induced Cramér-Rao kernel, i.e., the two variable function $K_{P_*}(z,w)$ defined in (14) yields a lower bound on the auto-covariance of any unbiased function estimator $\hat{P}$ of $P_* = \Pi(\theta^*)$.

In this section we will introduce the information metric and study its relation to the Cramér-Rao kernel. The main result of this section is that the Cramér-Rao kernel has the reproducing property (cf. Appendix A) with respect to the information metric.

In order to highlight the geometric aspects which are inherent in our setup, especially to the fact that the model set $\mathcal{P}$ is a differentiable manifold, we shall now adapt the standard information geometry terminology to suit our needs for function spaces.

Definition 3.1. For $\theta^* \in \Theta$ we define the *tangent space of* $\mathcal{P}$ at $P_* = \Pi(\theta^*)$ via

$$T_{P_*}\mathcal{P} := \{\Delta : \mathbb{T} \to \mathbf{C} \, | \, \exists \delta \in \mathbf{R}^d \text{ such that}$$
$$\forall z \in \mathbb{T} : \Delta(z) = \sum_{i=1}^{d} \delta_i \frac{\partial \Pi_z(\theta^*)}{\partial \theta_i} \}. \quad (16)$$

The J_{θ^*}-*induced information metric* on $\mathcal{P}$ at P_* is the inner-product $(\cdot, \cdot)_{P_*} : T_{P_*}\mathcal{P} \times T_{P_*}\mathcal{P} \to \mathbb{R}$ defined by

$$\left(\Delta^{(1)}, \Delta^{(2)} \right)_{P_*} := \delta^{(2),\mathrm{T}} J_{\theta^*} \delta^{(1)}, \quad (17)$$

for all $\Delta^{(l)}(z) = \sum_{i=1}^{d} \delta_i^{(l)} \frac{\partial \Pi_z(\theta^*)}{\partial \theta_i}$ with J_{θ^*} as in (13). $\blacksquare$

In Theorem 3.1 we shall prove the reproducing property of the J_{θ^*}-induced Cramér-Rao kernel with respect to the J_{θ^*}-induced information metric.

Theorem 3.1. Let $w \in \mathbb{T}$ be fixed. Then there exist uniquely defined functions $U, V \in T_{P_*}\mathcal{P}$ in the tangent space such that for all $\Delta \in T_{P_*}\mathcal{P}$ there holds

$$\Delta(w) = (\Delta, U)_{P_*} - \jmath \cdot (\Delta, V)_{P_*}. \quad (18)$$

The Cramér-Rao kernel is then *uniquely determined* by

$$K_{P_*}(z,w) = U(z) + j \cdot V(z) \quad \text{for all} \quad z \in \mathbb{T}. \quad (19)$$

We call this the *reproducing property of* K_{P_*} with respect to the J_{θ^*}-induced information metric.

Proof. The fact that U, V are uniquely defined follows from the Riesz-representation theorem for Hilbert spaces. Let $\Delta_i(z) = \frac{\partial \Pi_z(\theta^*)}{\partial \theta_i}$ for all $z \in \mathbb{T}$ and $i = 1, \ldots, d$.

To check (19) we note that

$$K_{P_*}(z,w) = \sum_{i,j=1}^{d} [J_{\theta_*}^{-1}]_{ji} \, \Delta_i(z) \cdot \overline{\Delta_j(w)},$$

and thus

$$K_{P_*}(z,w) = R(z) + \jmath \cdot J(z) \quad (20)$$

with

$$\begin{cases} R(z) = \sum_{i,j=1}^{d} [J_{\theta_*}^{-1}]_{ji} \, \Delta_i(z) \cdot \mathrm{Re}(\overline{\Delta_j(w)}), \\ J(z) = \sum_{i,j=1}^{d} [J_{\theta_*}^{-1}]_{ji} \, \Delta_i(z) \cdot \mathrm{Im}(\overline{\Delta_j(w)}), \end{cases}$$

for all $z \in \mathbb{T}$.

Let $\Delta \in T_{P_*}\mathcal{P}$ denote a tangent-vector. We expand

$$\Delta = \sum_{i=1}^{d} \delta_i \cdot \Delta_i \quad \text{with} \quad \delta \in \mathbf{R}^d,$$

and compute

$$\Delta(w) = [\Delta_1(w), \cdots, \Delta_d(w)]\delta$$
$$= [\mathrm{Re}\,\Delta_1(w), \cdots, \mathrm{Re}\,\Delta_d(w)]J_{\theta^*}^{-1} J_{\theta^*}\delta +$$
$$+ \jmath[\mathrm{Im}\,\Delta_1(w), \cdots, \mathrm{Im}\,\Delta_d(w)]J_{\theta^*}^{-1} J_{\theta^*}\delta$$
$$= (\Delta, R) - \jmath(\Delta, J).$$

By the uniqueness of U, V it follows that $R = U$ and $J = V$. Together with equation (20) this implies that indeed (19) holds. $\square$

4. REPARAMETRIZATION INVARIANCE

In this section we shall show that the lower bound K_{P_*} defined in (14), which we obtained in Lemma 2.1, is independent of the chosen parametrization $\Pi : \Theta \to \mathcal{P}$.

For this let $\Xi \subseteq \mathbb{R}^d$ denote an open subset such that

$$\mathcal{D} = \{\tilde{p}_\xi \,|\, \xi \in \Xi\} \quad \text{and} \quad \mathcal{P} = \{\tilde{\Pi}(\xi) \,|\, \xi \in \Xi\}, \quad (21)$$

denote alternative parametrizations of the set of densities and model set defined by (7) and (8), respectively, which obey the regularity assumptions R1) to R3). By assumption R1) there exists a unique smooth map $\varphi : \Xi \to \Theta$, the chart-transition, such that

$$\tilde{\Pi}(\xi) = \Pi(\varphi(\xi)) \quad \text{for all} \quad \xi \in \Xi. \quad (22)$$

In order to proceed we have to make a probabilistic compatibility assumption

$$\tilde{p}_\xi = p_{\varphi(\xi)} \quad \text{for all} \quad \xi \in \Xi. \quad (\tilde{R}1)$$

such that the abuse of notation in Remark 2.1 remains admissible.

Lemma 4.1. For $\xi^* \in \Xi$ and $P_* = \tilde{\Pi}(\xi^*)$ define

$$[\tilde{J}_{\xi^*}]_{ij} = \mathbb{E}_{\tilde{p}_{\xi^*}}\left[\frac{\partial \log \tilde{p}_{\xi^*}(\mathrm{x})}{\partial \xi_j} \cdot \frac{\partial \log \tilde{p}_{\xi^*}(\mathrm{x})}{\partial \xi_i}\right]. \quad (23)$$

Let $\theta^* \in \Theta$ such $\Pi(\theta^*) = P_*$ then

- There holds that

$$K_{P_*}(z,w) = \sum_{i,j=1}^{d}\left[\tilde{J}_{\xi^*}^{-1}\right]_{ji} \cdot \frac{\partial \tilde{\Pi}_z(\xi^*)}{\partial \xi_i}\, \overline{\frac{\partial \tilde{\Pi}_w(\xi^*)}{\partial \xi_j}}, \quad (24)$$

  i.e., the J_{ξ^*}-induced Cramér-Rao kernel (the right hand side of (24)) and J_{θ^*}-induced Cramér-Rao kernel defined in (14) are equal.

- The tangent-space (16) can be spanned using $\tilde{\Pi}$ instead of Π, i.e.,

$$T_{P_*}\mathcal{P} = \{\Delta : \mathbb{T} \to \mathbf{C}\,|\,\exists \delta \in \mathbf{R}^d \text{ such that}$$
$$\forall z \in \mathbb{T} : \Delta(z) = \sum_{i=1}^{d}\delta_i \frac{\partial \tilde{\Pi}_z(\theta^*)}{\partial \xi_i}\}.$$

- For $\Delta^{(l)}(z) = \sum_{i=1}^{d}\delta_i^{(l)}\frac{\partial \tilde{\Pi}_z(\xi^*)}{\partial \xi_i}$ with $l = 1, 2$, there holds

$$(\Delta^{(1)}, \Delta^{(2)})_{P_*} = \delta^{2,\mathrm{T}}\tilde{J}_{\xi^*}\delta^{(1)}, \quad (25)$$

  i.e., the J_{θ^*}-induced information metric (17) and the J_{ξ^*}-induced information metric (23) are equal.

Proof. Let $\varphi = (\varphi_1, \ldots, \varphi_d)$ and $L \in \mathbf{R}^{d \times d}$ denote the Jacobian defined by

$$[L]_{ij} = \frac{\partial \varphi_i(\xi^*)}{\partial \xi_j} \quad \text{for all} \quad i,j = 1, \ldots, d.$$

By the chain rule of differentiation from $(\tilde{R}1)$ and (22) it follows that

$$\frac{\partial \tilde{p}_{\xi^*}(x)}{\partial \xi_i} = \sum_{j=1}^{d}[L]_{ij} \cdot \frac{\partial p_{\theta^*}(x)}{\partial \theta_j} \quad \forall x \in X.$$

$$\frac{\partial \tilde{\Pi}_z(\xi^*)}{\partial \xi_i} = \sum_{j=1}^{d}[L]_{ij} \cdot \frac{\partial \Pi_z(\theta^*)}{\partial \theta_j} \quad \forall z \in \mathbb{T}.$$

This implies $\tilde{J}_{\xi^*} = L^{\mathrm{T}}J_{\theta^*}L$ and thus $J_{\theta^*}^{-1} = L\tilde{J}_{\xi^*}^{-1}L^{\mathrm{T}}$. In particular we have

$$K_{P_*}(z,w) = \sum_{i',j'=1}^{d}\left[J_{\theta^*}^{-1}\right]_{j'i'}\frac{\partial \Pi_z(\theta^*)}{\partial \theta_{i'}}\overline{\frac{\partial \Pi_w(\theta^*)}{\partial \theta_{j'}}}$$

$$= \sum_{i,i',j,j'=1}^{d}L_{j,j'}L_{i,i'}\left[\tilde{J}_{\xi^*}^{-1}\right]_{j,i} \cdot \frac{\partial \Pi_z(\theta^*)}{\partial \theta_{i'}}\overline{\frac{\partial \Pi_w(\theta^*)}{\partial \theta_{j'}}}$$

$$= \sum_{i,j}^{d}\left[\tilde{J}_{\xi^*}^{-1}\right]_{j,i} \cdot \frac{\partial \Pi_z(\xi^*)}{\partial \theta_i}\overline{\frac{\partial \Pi_w(\xi^*)}{\partial \xi_j}}$$

The remaining claims are simple consequences of the chain-rule of differentiation. This concludes the proof. □

5. COORDINATE-FREE DEFINITIONS

We recall that in Section 2 and 3 we defined the Cramér-Rao kernel K_{P_*} in (14), the tangent-space $T_{P_*}\mathcal{P}$ in (16), and the information metric $(\cdot, \cdot)_{P_*}$ in (17) w.r.t. a fixed parametrization. In Section 4 we proved that the resulting definitions are invariant if one uses another parametrization $\tilde{\Pi} : \Xi \to \mathcal{P}$ instead of Π. Our goal in this section will be to define the Cramér-Rao kernel and the information metric without invoking a parametrization.

It is a well known fact from differential geometry that there are various ways to define the tangent space to a differentiable manifold in a coordinate free way Lang (2002). One possibility is the following: For $P_* \in \mathcal{P}$ one defines the tangent space of $\mathcal{P}$ at P_* via

$$T_{P_*}\mathcal{P} := \{\Delta_\gamma : \mathbb{T} \to \mathbf{C}\,|\,\gamma : (-1,1) \to \mathcal{P}, \gamma(0) = P_*,$$
$$\forall z \in \mathbb{T} : \Delta_\gamma(z) = \frac{\mathrm{d}}{\mathrm{d}\tau}\mathrm{ev}_z(\gamma(\tau))_{|\tau=0}\}, \quad (26)$$

where $\mathrm{ev}_z : \mathcal{P} \to \mathbf{C}, P \to P(z)$ is smooth by R2), and the curve γ is always assumed smooth. This definition is equivalent to the one given by (16).

Definition 5.1. Let $\mu : \mathcal{P} \to \mathcal{D}, P \mapsto \mu_P$ denote a fixed smooth map. With respect to μ one defines the information metric via

$$(\Delta_{\gamma_1}, \Delta_{\gamma_2})_{P_*} :=$$
$$\mathbb{E}_{\mu_{P_*}}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\tau}\log \mu_{\gamma_2}(\mathrm{x})\big|_{\tau=0}\right)\left(\frac{\mathrm{d}}{\mathrm{d}\tau}\log \mu_{\gamma_1}(\mathrm{x})\big|_{\tau=0}\right)\right], \quad (27)$$

for all $\Delta_{\gamma_i} \in T_{P_*}\mathcal{P}$. We assume that the information metric is non-singular. [3]

For each fixed $w \in \mathbb{T}$ there exist uniquely defined functions $U_w, V_w \in T_{P_*}\mathcal{P}$ such that

$$\begin{cases} \mathrm{Re}\,\Delta(w) = (\Delta, U_w)_{P_*} \\ \mathrm{Im}\,\Delta(w) = -(\Delta, V_w)_{P_*}, \end{cases} \quad \text{for all} \quad \Delta \in T_{P_*}\mathcal{P}. \quad (28)$$

The function defined by

$$K_{P_*}(z,w) := U_w(z) + \jmath V_w(z) \quad \text{for all} \quad z,w \in \mathbb{T}, \quad (29)$$

is called the Cramér-Rao kernel. ∎

In Theorem 5.1 we establish that the nomenclature in Definition 5.1 is consistent with the earlier definitions made in Section 2 and Section 3.

Theorem 5.1. Assume that $\mu : \mathcal{P} \to \mathcal{D}$ is such that the probabilistic compatibility assumption

$$\mu_{\Pi(\theta)} = p_\theta \quad \text{for all} \quad \theta \in \Theta, \quad (\tilde{\tilde{R}}1)$$

is satisfied. For $\Pi_* = \Pi(\theta^*)$ the definitions of the information metric $(\cdot, \cdot)_{P_*}$ in (17) and (27) and the Cramér-Rao kernel K_{P_*} in (14) and (29) are equivalent.

Proof. Let $e_i = (0, \ldots, 0, \underbrace{1}_{i\text{-th pos.}}, 0, \ldots, 0) \in \mathbf{R}^d$ and

$$\gamma_i : (-1, 1) \to \mathcal{P}, \quad \text{with} \quad \gamma_i(\tau) = \Pi(\theta_* + \tau \cdot e_i).$$

Then one computes

$$[J_{\theta^*}]_{ij} = \mathbb{E}_{\mu_{P_*}}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\tau}\log \mu_{\gamma_j}(\mathrm{x})\big|_{\tau=0}\right)\left(\frac{\mathrm{d}}{\mathrm{d}\tau}\log \mu_{\gamma_i}(\mathrm{x})\big|_{\tau=0}\right)\right]$$

using the compatibility assumption $(\tilde{\tilde{R}}1)$. The equivalence of (14) and (29) then follows from the reproducing property proved in Theorem 3.1. □

6. APPLICATION TO SYSTEM IDENTIFICATION

In an important contribution in Ninness and Hjalmarsson (2004) it was shown that the asymptotic average variability of a transfer-function estimator at a fixed frequency $z \in \mathbb{T}$ is given by $K_{P_*}(z,z)$, i.e., the Cramér-Rao kernel evaluated on its diagonal.

---

[3] The metric $(\cdot, \cdot)_{P_*}$ is called non-singular if $(\Delta, \Delta)_{P_*} = 0$ implies that $\Delta = 0$ for all $\Delta \in T_{P_*}\mathcal{P}$.

The goal of this section is twofold. In Section 6.1 we demonstrate how the asymptotic average Cramér-Rao kernel can be computed without invoking a parametrization. In Section 6.2 we stress that the off-diagonal elements of the Cramér-Rao kernel, i.e., $K_{P_*}(z, w)$ with $z \neq w$, are equally important as the diagonal elements.

6.1 Computation of the Cramér-Rao Kernel

In the preceding sections the information metric quantified the "information" contained in one sample $x_1 \in X$ where $X \subseteq \mathbf{R}^q$ denotes the sample space. In applications like System Identification an observation consists of a sample vector

$$x^N = (x_1, \ldots, x_N) \in X^N, \tag{30}$$

where N is called the sample-size. The goal of this section will be to quantify the asymptotic average Cramér-Rao kernel, i.e., to quantify the autocovariance of an efficient transfer function estimator.

If one replaces x in (27) by the sample vector x^N given by (30) one obtains a *sample size dependent* information metric. To distinguish between different sample sizes we use the notation $(\cdot, \cdot)_{N, P_*}$ to denote the information-metric given N samples.

As an example assume $\mathcal{P}$ denotes a model set consisting of asymptotically stable transfer-functions, and

$$x_t = (u_t, y_t) \in X = \mathbf{R}^2 \quad \text{s.t.} \quad y = P_* u + v, \tag{31}$$

where the noise-process v and the input-process u are assumed independent, with spectra given by $\Phi_v, \Phi_u : \mathbb{T} \to \mathbf{R}$ respectively. [4]

A fundamental contribution in Caines and Ljung (1979) was to show that, with sample size $N \to \infty$, the expression $N^{-1}(\Delta_1, \Delta_2)_{N, P_*}$ tends, for all $\Delta_i \in T_{P_*}\mathcal{P}$, to a signal-to-noise ratio weighted L_2-inner-product

$$\langle \Delta_1, \Delta_2 \rangle_{P_*} := \frac{1}{2\pi} \int_{-\pi}^{\pi} \Delta_1(e^{j\omega}) \overline{\Delta_2(e^{j\omega})} \frac{\Phi_u(e^{j\omega})}{\Phi_v(e^{j\omega})} \, d\omega, \tag{32}$$

which is therefore called the metric which measures the *average information per sample*.

The next theorem shows that there exist cases where it is possible to compute the tangent-space and the Cramér-Rao kernel without using a parametrization. The key advantage of our geometric point of view is that: it allows the usage of existing knowledge on tangent spaces of transfer-function manifolds Helmke and Fuhrmann (1998) and structured reproducing kernel Hilbert spaces Grenander and Szego (1958). In other words, the coordinate free approach renders it unnecessary to rederive such results in a parametric setup.

Theorem 6.1. Let $\mathrm{Rat}(n)$ denote the $2n$-dimensional manifold of all i) real rational ii) strictly proper, iii) Schur stable transfer functions of McMillan-degree n, i.e.,

---

$$\mathrm{Rat}(n) := \left\{ \frac{b}{p} \,\middle|\, \text{i) } b, p \in \mathbb{R}[z] \text{ are coprime s.t.} \right.$$
$$\text{ii) } \deg(b) < \deg(p) = n, \text{ and} \tag{33}$$
$$\left. \text{iii) } p(z) = 0 \quad \text{implies} \quad |z| < 1 \right\}.$$

The tangent-space at $b/p \in \mathrm{Rat}(n)$ is given by

$$T_{b/p}\mathrm{Rat}(n) := \mathrm{X}^{p^2}, \tag{34}$$

where for any non-zero polynomial q in $\mathbb{R}[z]$ we define

$$\mathrm{X}^q := \left\{ \frac{f}{q} \,\middle|\, f \in \mathbb{R}[z] \text{ with } \deg(f) < \deg(q) \right\}. \tag{35}$$

For $\Phi_u/\Phi_v = 1$ the Cramér-Rao kernel w.r.t. to the asymptotic average information, is given by

$$\kappa_{b/p}(z, w) = \frac{\left(\frac{z^n p(1/z)}{p(z)} \cdot \frac{p(w)}{w^n p(1/w)} \right)^2 - 1}{1 - z/w}, \tag{36}$$

for all $z, w \in \mathbb{T}$. For $\Phi_u/\Phi_v \equiv \rho$ for some constant ρ the asymptotic Cramér-Rao kernel is $K_{b/p} = \kappa_{b/p}/\rho$ which, when evaluated on the diagonal, (i.e., for $z = w$), yields

$$K_{b/p}(z, z) = \frac{2\Phi_v}{\Phi_u} \sum_{i=1}^{n} \frac{1 - |a_i|^2}{|z - a_i|^2}, \tag{37}$$

where $p(z) = \prod_{i=1}^{n}(z - a_i)$.

Proof. The fact that (34) holds has been shown in Helmke and Fuhrmann (1998). The fact that (36) is the reproducing kernel of the complexification of X^{p^2}, i.e., of the space

$$\mathbf{X}^{p^2} = \{ \frac{q}{p^2} \,|\, q \in \mathbf{C}[s], \deg(q) < 2n \}, \tag{38}$$

is a well known fact, see e.g. Ninness and Gustafsson (1994). By Remark A.1 in Appendix A this implies the reproducing property, i.e., that $\kappa_{b/p}$ defined in (36) is indeed the asymptotic average Cramér-Rao kernel. The remaining equation (37) follows from (36) by straightforward calculation. $\square$

Remark 6.1. Our assumption that the signal to noise ratio ρ is constant leads to insight as to how the poles of the true system influence the Cramér-Rao kernel. In the general case this dependence is more complicated. Nevertheless one can obtain the Cramér-Rao kernel for the case $\Phi_u/\Phi_v = \rho \not\equiv$ const using Szego's formula (Grenander and Szego, 1958, §2.3):

$$K_{b/p}(z, w) = \frac{1/\alpha}{p(z)p(\bar{w})} \cdot \frac{\varphi(-z)\varphi(-\bar{w}) - z\bar{w}\varphi(z)\varphi(w)}{1 - z\bar{w}}, \tag{39}$$

for all $z, w \in \mathbb{T}$, where

$$\alpha = \frac{1}{2\pi} \int_{-\pi}^{\pi} |\varphi(e^{j\omega})|^2 \frac{\rho(e^{j\omega})}{|p(e^{j\omega})|^2} d\omega, \tag{40}$$

and

$$\varphi(z) = \begin{vmatrix} c_0 & \cdot & \cdot & c_{-2n+1} & c_{-2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ c_{2n-1} & \cdot & \cdot & c_0 & c_{-1} \\ 1 & \cdot & \cdot & z^{2n-1} & z^{2n} \end{vmatrix}, \tag{41}$$

and $c_{-2n}, \ldots, c_{2n}$ denote the generalized moments of the signal to noise ratio, i.e.,

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{j \cdot k \, \omega} \cdot \frac{\rho(e^{j\omega})}{|p(e^{j\omega})|^2} d\omega \quad \text{for} \quad k = \pm 1, \ldots, \pm 2n.$$

---

[4] One assumes that $\Phi_v(z) = \sigma^2 |H(z)|^2$ for a stable, minimum-phase transfer function H. Moreover, due to the stability assumption on P, for large sample size N, equation (31) can be assumed to have zero initial conditions.

6.2 Autocovariance versus classical variance function

It is a natural question to ask whether the correlation between the mismatch at different frequencies can be neglected. In other words, does the diagonal of the Cramér-Rao kernel given in (37) capture the behavior of the estimator. Unfortunately this is not the case, i.e., the properties of an estimator can change dramatically if the autocovariance changes, even if the variance does not change. For the rest of this section we shall assume that we are given two transfer function estimators which collect samples under the same experimental conditions (31); however, they differ in the number of samples they collect and the model structure they use, though the true plant P_* lies in both model sets. . To be concrete we assume that

- $\hat{P}$ is given the sample vector $x^N = (x_1, \ldots, x_N)$ and yields a model in $\mathcal{P}$,
- $\tilde{P}$ is given the sample vector $x^{\tilde{N}} = (x_1, \ldots, x_{\tilde{N}})$ and yields a model in $\tilde{\mathcal{P}}$,

where both sample sizes $N, \tilde{N}$ are assumed to be large for asymptotic formulas to be applicable.

Let K_{P_*} and $\tilde{K}_{P_*}$ denote the asymptotic average Cramér-Rao kernels corresponding to the model sets $\mathcal{P}$ and $\tilde{\mathcal{P}}$, respectively. It is clear that

$$\frac{K_{P_*}}{N} \leq_{\text{EH}} \frac{\tilde{K}_{P_*}}{\tilde{N}} \quad \Rightarrow \quad \forall z \in \mathbb{T}: \ \frac{K_{P_*}(z,z)}{N} \leq \frac{\tilde{K}_{P_*}(z,z)}{\tilde{N}}. \tag{42}$$

However, the converse need not to be true. That is to say it may very well be that the right hand side of (42) holds while the left-hand-side is false. From a statistical point of view this means that given a scalar $g(P_*) \in \mathbf{C}$ function of the plant $P_* \in \mathcal{P} \cap \tilde{\mathcal{P}}$ the right-hand-side of (42), i.e., the variance function of $\hat{P}_{x^N}$ being smaller than the variance function of $\tilde{P}_{x^{\tilde{N}}}$, does **not** imply that

$$\mathbb{E}_{P_*}[|g(\hat{P}_{x^N}) - g(P_*)|^2] \leq \mathbb{E}_{P_*}[|g(\tilde{P}_{x^{\tilde{N}}}) - g(P_*)|^2]. \tag{43}$$

The next example shows that this fact is not as counter-intuitive as it might seem at first.

Example 6.1. Let FIR(n) denote the model set consisting of finite impulse response filters of order n. Let

$$\mathcal{P} = \text{FIR}(4), \ \tilde{\mathcal{P}} = \text{FIR}(40), \quad \text{and} \quad P_* \in \mathcal{P} \cap \tilde{\mathcal{P}}.$$

The goal is to estimate $g(P_*)$ where $g(\cdot) = \|\cdot\|_2^2$ denotes the squared standard L_2-norm. For simplicity assume the samples are collected from (31) with $\Phi_u = \Phi_v \equiv 1$.

By straightforward calculation one obtains

$$K_{P_*}(z,z) \equiv 4 \quad \text{and} \quad \tilde{K}_{P_*}(z,z) \equiv 40,$$

for all $z \in \mathbb{T}$. For the sample-size configuration $\tilde{N} = 2 \cdot N$ this means that

$$5 \cdot \mathbb{E}_{P_*}[|\hat{P}_{x^N}(z) - P_*(z)|^2] = \mathbb{E}_{P_*}[|P_{x^{\tilde{N}}}(z) - P_*(z)|^2],$$

for all $z \in \mathbb{T}$, i.e., the variance function of $\hat{P}_{x^N}$ is 5-times smaller than the variance function of $\tilde{P}_{x^{\tilde{N}}}$.

However, by analysis or by means of Monte-Carlo simulations, one verifies that

$$\mathbb{E}_{P_*}[|g(\hat{P}_{x^N}) - g(P_*)|^2] = 2 \cdot \mathbb{E}_{P_*}[|g(\tilde{P}_{x^{\tilde{N}}}) - g(P_*)|^2],$$

i.e., the variance of the estimator $g(\hat{P}_{x^N})$ is two times larger than the variance of $g(\tilde{P}_{x^{\tilde{N}}})$.
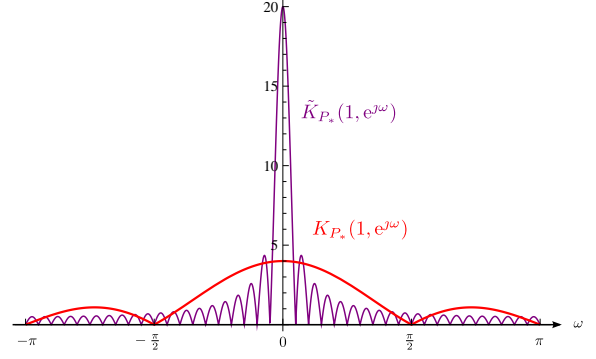

Fig. 1. The Cramér-Rao kernels K_{P_*} and $\tilde{K}_{P_*}$ are equal to the autocovariance functions of $\hat{P}$ and $\tilde{P}$ respectively when normalized by sample size N. Note that both processes are stationary in the sense that

$$\text{Cov}_{P_*}(\hat{P})(z,w) = \text{Cov}_{P_*}(\hat{P})(1, z^{-1}w),$$

for all $z, w \in \mathbb{T}$ and similarly for $\tilde{P}$. The graph shows that the correlation of the mismatch of the high order estimator $\tilde{P}$ at different frequencies decays a lot faster than for low order estimator $\hat{P}$.

Intuition tells us that the mismatch of the low order estimator $\hat{P}_{x^N} - P_*$ at one frequency is correlated to the mismatch at another frequency. This is due to the constraints imposed by the model set $\mathcal{P}$. These correlations are far less pronounced for the mismatch of the high order estimator $\tilde{P}_{x^N} - P_*$. This is due to the flexibility offered by the high dimensional model set $\tilde{\mathcal{P}}$. Both phenomena are reflected by the off-diagonal elements of the Cramér-Rao kernel; see Fig. 1. Since g takes an average over all frequencies, correlations of the mismatch are harmful, i.e., the less correlation the better. It is therefore intuitively not too surprising that $g(\tilde{P}_{x^{\tilde{N}}})$ with $\tilde{N} = 2N$ samples is the better estimator for the squared L_2-norm of the system. It has twice more samples than $g(\hat{P}_{x^N})$. [5] ∎

7. ROBUST PERFORMANCE SPECIFICATIONS

In this section we want to demonstrate how the autocovariance can be used for input design aiming at robust performance specifications. Such performance specifications naturally arise if one performs application oriented identification Hjalmarsson (2009); Gevers (2005). In the following we shall derive a sufficient condition for an efficient estimator $\hat{P}$ of $P_* \in \mathcal{P}$ to satisfy the following performance specification:

$$\alpha \leq \text{Prob}\{x \in X^N : \sup_{z \in B} |\hat{P}_x(z) - P_*(z)| \leq \frac{\varepsilon}{N}\}. \tag{44}$$

In other words the experimental conditions satisfy the performance specification if, with a probability not less than α, we have x is a good sample, i.e., that the supremum norm of the mismatch $\hat{P}_x - P$ on the frequency band $B \subseteq \mathbb{T}$ is bounded from above by ε/N.

Theorem 7.1. An asymptotically (in N) sufficient condition for (44) is given by

---

[5] We note that if $\tilde{N} = N$, i.e., if the sample sizes would have been equal, both estimators would have been equally good.

$$\forall z \in B: \quad K_{P_*}(z,z) + |K_{P_*}(z,z^{-1})| \leq \frac{2 \cdot \varepsilon}{\chi_\alpha^2(d)}, \quad (45)$$

where:

- d is the dimension of the differentiable manifold $\mathcal{P}$,
- $\chi_\alpha^2(d)$ is the α-quantile of the Chi-square distribution with d degrees of freedom,
- K_{P_*} is the asymptotic average Cramér-Rao kernel.

Proof. As $N \to \infty$ the mismatch $\hat{P}_{\mathrm{x}} - P$ can be approximated with a random, i.e., sample dependent, tangent vector $\Delta_{\mathrm{x}} \in T_P\mathcal{P}$ which, by asymptotic normality, can be assumed Gaussian with zero mean and Chi-square distributed squared norm, i.e.,

$$\|\Delta_{\mathrm{x}}\|_{P_*}^2 := \langle \Delta_{\mathrm{x}}, \Delta_{\mathrm{x}} \rangle \sim \chi^2(d).$$

The α-quantile, denoted by $r := \chi_\alpha^2(d)$, is defined as the inverse of the cumulative distribution of $\|\Delta_{\mathrm{x}}\|_{P_*}^2$ evaluated at α. This is equivalent to

$$\mathrm{Prob}(B_r) = \alpha \quad \text{with} \quad B_r = \{x \in X \,|\, \|\Delta_x\|_{P_*}^2 \leq r\}.$$

A result established in Ivanov et al. (2009) guarantees that

$$2 \cdot \Delta_x(z) \leq \|\Delta_x\|_{P_*}^2 \cdot (K_{P_*}(z,z) + |K_{P_*}(z,z^{-1})|),$$

holds for all $x \in X$ and all $z \in \mathbb{T}$. So condition (45) implies that the event

$$\tilde{B}_\varepsilon = \{x \in X \,|\, \forall z \in B : \Delta_x(z) \leq \varepsilon\},$$

is implied by the event $\mathrm{x} \in B_r$. From this it follows that $\tilde{B}_\varepsilon$ has a probability greater than or equal to α. $\quad\square$

Before we conclude we give an example of the structural insight which can be obtained from Theorem 7.1.

Example 7.1. Let $P \in \mathcal{P}$ with $\mathcal{P} = \mathrm{Rat}(2)$ denote a plant with a complex conjugate pole at $a \in \mathbb{C}$ such that

$$P = b/p \quad \text{where} \quad p(z) = (z-a)(z-a^*),$$

where $|a| < 1$ and $b, p \in \mathbb{R}[z]$ coprime. Moreover let

$$\alpha = 0.9, \ B = \{e^{j\omega} \,|\, |\omega| < \beta\} \quad \text{and} \quad \frac{\varepsilon}{100} = 0.1.$$

Assume we want to meet the performance specification (44), under the assumption $\Phi_v \equiv \mathrm{const}$ and $\Phi_u \equiv \mathrm{const}$. The dependence of the minimum amount of input power required, $\Phi_u \geq \Phi_{u,\mathrm{min}}$ depends on the bandwidth $\beta \in (0,\pi)$ and the pole location $a \in \mathbb{C}$. This is illustrated in Fig. 2 for constant absolute value and varying phase of a and in Fig. 3 for constant phase and varying absolute value. Let

$$C_P(z) = \frac{1}{2}\kappa_P(z,z) + \frac{1}{2}|\kappa_P(z,z^{-1})|,$$

and κ_P is given by (36). By Theorem 6.1

$$\Phi_v \cdot \chi_{0.9}^2(4) \cdot C_P(e^{j\omega}) \leq \varepsilon \cdot \Phi_u \quad \text{for all} \quad \omega \in [0,\beta].$$

is equivalent to the inequality (45). $\quad\blacksquare$

8. CONCLUSIONS

In this paper we have established that the autocovariance of an unbiased function estimator is a positive kernel which can be bounded from below by the reproducing kernel of the tangent space of the model structure. Therefore, in the context of prediction error identification of linear time invariant systems, the quantification of the autocovariance of a transfer function estimator can be split into two
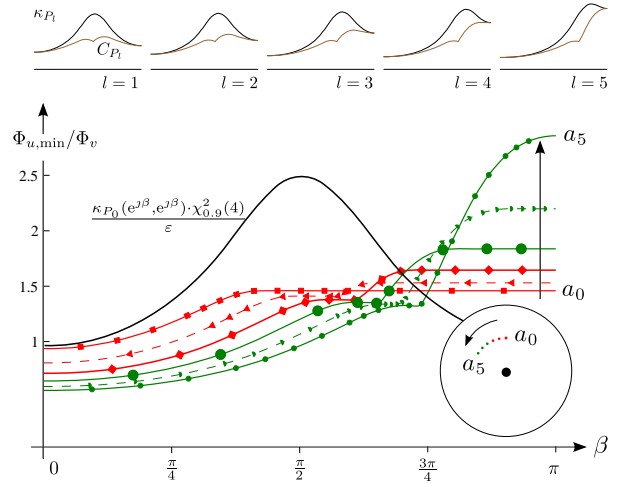
Fig. 2. Minimum input power $\Phi_{u,\mathrm{min}}(\beta)$ as a function of bandwidth β for six different plants P_l with $l = 0, \ldots, 5$ with a complex conjugate pole $a_l = \frac{1}{2}e^{j(\pi/2 + l/5)}$ respectively. At the top: Reproducing kernel functions $\kappa_{P_l}(z,w)$ with respect to standard L^2-inner-product, evaluated on the diagonal elements $z = w = e^{j\omega}$, with ω varying in $[0,\pi]$, as well as the function $C_{P_l}(z) = \frac{1}{2}\kappa_{P_l}(z,z) + \frac{1}{2}|\kappa_{P_l}(z,z^{-1})|$ for $z = e^{j\omega}$ At the bottom: Minimal amount of input power $\Phi_{u,\mathrm{min},l}(\beta)$ needed to achieve the desired accuracy for $P = P_l$ on the frequency band $[-\beta,\beta]$, i.e, to turn the inequality (7.1) into an equality.
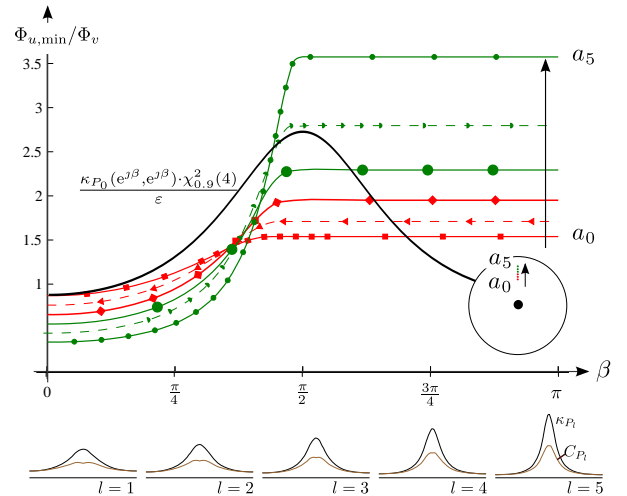

Fig. 3. Same setup as in Figure 2. However this time each plant P_l has a complex conjugate pole at $a_l = j(4/5 - 4(5-l)/75)$ with $l = 0, \ldots, 5$, respectively.

subproblems: determining the tangent space of the model manifold at the system which generated the data, and computing its reproducing kernel with respect to a signal to noise ratio weighted L_2-inner-product. We have given an example of how to handle such performance specifications for the case where the specification required the supremum norm of the mismatch between estimator and explanatory model to be less than some constant with prescribed probability.

REFERENCES

Amari, S.I. and Nagaoka, H. (2001). *Methods of Information Geometry (Translations of Mathematical Monographs)*, volume 191. American Mathematical Society.

Caines, P. and Ljung, L. (1979). Asymptotic normality of prediction error estimators for approximate system models. *Stochastics An International Journal of Probability and Stochastic Processes*, 3, 29–46.

Gevers, M. (2005). Identification for control: from the early achievements to the revival of experiment design. *semi-plenary lecture at CDC-ECC 2005, European Journal of Control*, 11, 1–18.

Gevers, M., Bazanella, A., and Bombois, X. (2009). Connecting informative experiments, the information matrix and the minima of a prediction error identification criterion. *Proc. of 15th IFAC Symposium on System Identification (CD-ROM)*, 675–680.

Grenander, U. and Szego, G. (1958). *Toeplitz Forms and Their Applications*. University of California Press, Berkley.

Helmke, U. and Fuhrmann, P. (1998). Tangent spaces of rational matrix functions. *Linear Algebra and its Applications*, 271(1-3), 1 – 40.

Hjalmarsson, H. (2009). System identification of complex structured systems. *Plenary lecture at ECC 2009, European Journal of Control*, 15(3-4), 276–310.

Ivanov, T., B.D.O. Anderson, P.-A. Absil, and Gevers, M. (2009). Using H^2 norm to bound H^∞ norm from above on Real Rational Modules. In *Proc. of European Control Conference CD-ROM*, 2259–2264.

Lang, S. (2002). *Introduction to Differentiable Manifolds*. Springer-Verlags.

Moore, E.H. (1916). On properly positive Hermitian matrices. *Bull. Amer. Math. Soc.*, 23(59), 66–67.

Ninness, B. and Gustafsson, F. (1994). A unifying construction of orthonormal bases for system identification. In *IEEE Conference on Decision and Control*, 3388–3393.

Ninness, B. and Hjalmarsson, H. (2004). Variance error quantifications that are exact for finite-model order. *Transactions on Automatic Control*, 49(8), 1275–1290.

Rao, C.R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of Calcutta Mathematical Society*, 37, 81–91.

Appendix A. REPRODUCING KERNELS

Let H denote a linear space over the complex numbers with $\dim H = n < \infty$ endowed with an inner-product

$$\langle \cdot, \cdot \rangle : H \times H \to \mathbb{C}.$$

If the elements of H are functions defined on a common set Ω, then there exists a unique two variable function $K : \Omega \times \Omega \to \mathbf{C}$ such that

$$\langle f, K_w \rangle = f(w) \quad \text{with} \quad K_w(z) = K(z,w), \qquad (A.1)$$

holding for all $z, w \in \Omega$ and all $f \in H$.

One calls K the reproducing kernel of $(H, \langle \cdot, \cdot \rangle)$ and (A.1) the reproducing property. There are various ways to compute K given H and $\langle \cdot, \cdot \rangle$. A general method, known as the Aitken-Berg-Collar Lemma, is given by

$$K(z,w) = \sum_{i,j=1}^{n} [G^{-1}]_{ji} \, b_i(z) \overline{b_j(w)}, \qquad (A.2)$$

where $b_1, \ldots, b_n$ is an arbitrary basis of H and $G_{ij} = \langle b_j, b_i \rangle$ the associated Gramian.

Remark A.1. Let $b_1, \ldots, b_n$ denote a basis of H with $\langle b_j, b_i \rangle \in \mathbf{R}$ and let

$$X = \left\{ \sum_{i=1}^{n} \alpha_i \, b_i \,\middle|\, \alpha \in \mathbf{R}^n \right\}. \qquad (A.3)$$

Moreover define the bilinear inner-product via

$$(\cdot, \cdot) : X \times X \to \mathbf{R} \quad \text{with} \quad (f,g) = \langle f, g \rangle, \qquad (A.4)$$

for all $f, g \in X$. Then (A.1) is equivalent to

$$K(z,w) = U_w(z) + \jmath V_w(z), \qquad (A.5)$$

with $U_w, V_w \in X$ uniquely determined by

$$\begin{cases} \operatorname{Re} f(w) = (f, U_w), \\ \operatorname{Im} f(w) = -(f, V_w), \end{cases} \quad \text{for all} \quad f \in X. \qquad (A.6)$$