Copyright 2014 IEEE IEEE Trans. Inf. Theory, vol. 60, no. 7, July 2014, pp. 4050–4065

The RKHS Approach to Minimum Variance Estimation Revisited: Variance Bounds, Sufficient Statistics, and Exponential Families

Alexander Jung, *Member, IEEE*, Sebastian Schmutzhard, and Franz Hlawatsch, *Fellow, IEEE*

***Abstract*—The mathematical theory of reproducing kernel Hilbert spaces (RKHS) provides powerful tools for minimum variance estimation (MVE) problems. Here, we extend the classical RKHS-based analysis of MVE in several directions. We develop a geometric formulation of five known lower bounds on the estimator variance (Barankin bound, Cramér–Rao bound, constrained Cramér–Rao bound, Bhattacharyya bound, and Hammersley-Chapman-Robbins bound) in terms of orthogonal projections onto a subspace of the RKHS associated with a given MVE problem. We show that, under mild conditions, the Barankin bound (the tightest possible lower bound on the estimator variance) is a lower semi-continuous function of the parameter vector. We also show that the RKHS associated with an MVE problem remains unchanged if the observation is replaced by a sufficient statistic. Finally, for MVE problems conforming to an exponential family of distributions, we derive novel closed-form lower bounds on the estimator variance and show that a reduction of the parameter set leaves the minimum achievable variance unchanged.**

***Index Terms*—Minimum variance estimation, exponential family, reproducing kernel Hilbert space, RKHS, Cramér–Rao bound, Barankin bound, Hammersley–Chapman–Robbins bound, Bhattacharyya bound, locally minimum variance unbiased estimator.**

I. Introduction

We consider the problem of estimating the value $\mathbf{g}(\mathbf{x})$ of a known deterministic function $\mathbf{g}(\cdot)$ evaluated at an unknown nonrandom parameter vector $\mathbf{x} \in \mathcal{X}$, where the parameter set $\mathcal{X}$ is known. The estimation of $\mathbf{g}(\mathbf{x})$ is based on an observed vector $\mathbf{y}$, which is modeled as a random vector with an associated probability measure [1] $\mu_{\mathbf{x}}^{\mathbf{y}}$ or, as a special case, an associated probability density function (pdf) $f(\mathbf{y};\mathbf{x})$, both parametrized by $\mathbf{x} \in \mathcal{X}$. More specifically, we study the problem of *minimum variance estimation* (MVE), where one aims at finding estimators with minimum variance under the constraint of a prescribed bias. Our treatment of MVE will be based on the mathematical framework and methodology of *reproducing kernel Hilbert spaces* (RKHS).

A. State of the Art and Motivation

The RKHS approach to MVE was introduced in the seminal papers [2] and [3]. On a general level, the theory of RKHS yields efficient methods for high-dimensional optimization problems. These methods are popular, e.g., in machine learning [4], [5]. For the MVE problem considered here, the optimization problem is the minimization of the estimator variance subject to a bias constraint. The RKHS approach to MVE enables a consistent and intuitive geometric treatment of the MVE problem. In particular, the determination of the minimum achievable variance (or Barankin bound) and of the locally minimum variance estimator reduces to the calculation of the squared norm and isometric image of a specific vector—representing the prescribed estimator bias—that belongs to the RKHS associated with the estimation problem. This reformulation is interesting from a theoretical perspective; in addition, it may also be the basis for an efficient computational evaluation. Furthermore, a wide class of lower bounds on the minimum achievable variance (and, in turn, on the variance of any estimator) is obtained by performing projections onto subspaces of the RKHS. Again, this enables an efficient computational evaluation of these bounds.

A specialization to estimation problems involving sparsity constraints was presented in [6]–[8]. For certain special cases of these sparse estimation problems, the RKHS approach allows the derivation of closed-form expressions of the minimum achievable variance and the corresponding locally minimum variance estimators. The RKHS approach has also proven to be a valuable tool for the analysis of estimation problems involving continuous-time random processes [2], [3], [9].

B. Contribution and Outline

The main contributions of this paper concern an RKHS-theoretic analysis of the performance of MVE, with a focus on lower variance bounds, sufficient statistics, and observations conforming to an exponential family of distributions. First, we give a geometric interpretation of some well-known lower bounds on the estimator variance. The tightest of these bounds, i.e., the Barankin bound, is proven to be a lower semi-continuous function of the parameter vector $\mathbf{x}$ under mild conditions. We then analyze the role of a sufficient

A. Jung and F. Hlawatsch are with the Institute of Telecommunications, Vienna University of Technology, A-1040 Vienna, Austria (e-mail: {ajung,fhlawats}@nt.tuwien.ac.at). S. Schmutzhard is with the Numerical Harmonic Analysis Group, Faculty of Mathematics, University of Vienna, A-1090 Vienna, Austria (e-mail: sebastian.schmutzhard@univie.ac.at). This work was supported in part by the Austrian Science Fund under Grants S10602-N13 and S10603-N13 within the National Research Network Signal and Information Processing in Science and Engineering and in part by the Wiener Wissenschafts-, Forschungs- und Technologiefonds (WWTF) under Grant MA 07-004.

statistic from the RKHS viewpoint. In particular, we prove that the RKHS associated with an estimation problem remains unchanged if the observation $\mathbf{y}$ is replaced by any sufficient statistic. Furthermore, we characterize the RKHS for estimation problems with observations conforming to an exponential family of distributions. It is found that this RKHS has a strong structural property, and that it is explicitly related to the moment-generating function of the exponential family. Inspired by this relation, we derive novel lower bounds on the estimator variance, and we analyze the effect of parameter set reductions. The lower bounds have a particularly simple form.

The remainder of this paper is organized as follows. In Section II, basic elements of MVE are reviewed and the RKHS approach to MVE is summarized. In Section III, we present an RKHS-based geometric interpretation of known variance bounds and demonstrate the lower semi-continuity of the Barankin bound. The effect of replacing the observation by a sufficient statistic is studied in Section IV. In Section V, the RKHS for exponential family-based estimation problems is investigated, novel lower bounds on the estimator variance are derived, and the effect of a parameter set reduction is analyzed. We note that the proofs of most of the new results presented can be found in the doctoral dissertation [10] and will be referenced in each case.

C. Notation and Basic Definitions

We will use the shorthand notations $\mathbb{N} \triangleq \{1,2,3,\ldots\}$, $\mathbb{Z}_+ \triangleq \{0,1,2,\ldots\}$, and $[N] \triangleq \{1,2,\ldots,N\}$. The open ball in $\mathbb{R}^N$ with radius $r > 0$ and centered at $\mathbf{x}_c$ is defined as $\mathcal{B}(\mathbf{x}_c, r) \triangleq \{\mathbf{x} \in \mathbb{R}^N \,|\, \|\mathbf{x}-\mathbf{x}_c\|_2 < r\}$. We call $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^N$ an *interior point* if $\mathcal{B}(\mathbf{x}, r) \subseteq \mathcal{X}$ for some $r > 0$. The set of all interior points of $\mathcal{X}$ is called the *interior* of $\mathcal{X}$ and denoted $\mathcal{X}^{\mathrm{o}}$. A set $\mathcal{X}$ is called *open* if $\mathcal{X} = \mathcal{X}^{\mathrm{o}}$.

Boldface lowercase (uppercase) letters denote vectors (matrices). The superscript ^{T} stands for transposition. The kth entry of a vector $\mathbf{x}$ and the entry in the kth row and lth column of a matrix $\mathbf{A}$ are denoted by $(\mathbf{x})_k = x_k$ and $(\mathbf{A})_{k,l} = A_{k,l}$, respectively. The kth unit vector is denoted by $\mathbf{e}_k$, and the identity matrix of size $N \times N$ by $\mathbf{I}_N$. The Moore-Penrose pseudoinverse [11] of a rectangular matrix $\mathbf{A} \in \mathbb{R}^{M\times N}$ is denoted by $\mathbf{A}^\dagger$.

A function $f(\cdot)\colon \mathcal{D} \to \mathbb{R}$, with $\mathcal{D} \subseteq \mathbb{R}^N$, is said to be *lower semi-continuous* at $\mathbf{x}_0 \in \mathcal{D}$ if for every $\varepsilon > 0$ there is a radius $r > 0$ such that $f(\mathbf{x}) \geq f(\mathbf{x}_0) - \varepsilon$ for all $\mathbf{x} \in \mathcal{B}(\mathbf{x}_0, r)$. (This definition is equivalent to $\liminf_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) \geq f(\mathbf{x}_0)$, where $\liminf_{\mathbf{x}\to\mathbf{x}_0} f(\mathbf{x}) \triangleq \sup_{r>0} \big\{\inf_{\mathbf{x}\in\mathcal{D}\cap[\mathcal{B}(\mathbf{x}_0,r)\setminus\{\mathbf{x}_0\}]} f(\mathbf{x})\big\}$ [12], [13].) The restriction of a function $f(\cdot) : \mathcal{D} \to \mathbb{R}$ to a subdomain $\mathcal{D}' \subseteq \mathcal{D}$ is denoted by $f(\cdot)\big|_{\mathcal{D}'}$. Given a multi-index $\mathbf{p} = (p_1 \cdots p_N)^T \in \mathbb{Z}_+^N$, we define the partial derivative of order $\mathbf{p}$ of a real-valued function $f(\cdot) : \mathcal{D} \to \mathbb{R}$, with $\mathcal{D} \subseteq \mathbb{R}^N$, as $\frac{\partial^{\mathbf{p}} f(\mathbf{x})}{\partial \mathbf{x}^{\mathbf{p}}} \triangleq \frac{\partial^{p_1}}{\partial x_k^{p_1}} \cdots \frac{\partial^{p_N}}{\partial x_N^{p_N}} f(\mathbf{x})$ (if it exists) [13], [14]. Similarly, for a function $f(\cdot,\cdot) : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ and two multi-indices $\mathbf{p}_1, \mathbf{p}_2 \in \mathbb{Z}_+^N$, we denote by $\frac{\partial^{\mathbf{p}_1}\partial^{\mathbf{p}_2} f(\mathbf{x}_1,\mathbf{x}_2)}{\partial \mathbf{x}_1^{\mathbf{p}_1} \partial \mathbf{x}_2^{\mathbf{p}_2}}$ the partial derivative of order $(\mathbf{p}_1, \mathbf{p}_2)$, where $f(\mathbf{x}_1, \mathbf{x}_2)$ is considered as a function of the "super-vector" $(\mathbf{x}_1^T\, \mathbf{x}_2^T)^T$ of length $2N$. Given a vector-valued function $\boldsymbol{\phi}(\cdot) : \mathbb{R}^M \to \mathbb{R}^N$ and $\mathbf{p} \in \mathbb{Z}_+^N$, we denote the product $\prod_{k=1}^N \big(\phi_k(\mathbf{y})\big)^{p_k}$ by $\boldsymbol{\phi}^{\mathbf{p}}(\mathbf{y})$.

The probability measure of a random vector $\mathbf{y}$ taking on values in $\mathbb{R}^M$ is denoted by $\mu^{\mathbf{y}}$ [1], [15]–[17]. We consider probability measures that are defined on the measure space given by all M-dimensional Borel sets on $\mathbb{R}^M$ [1, Sec. 10]. The probability measure assigns to a measureable set $\mathcal{A} \subseteq \mathbb{R}^M$ the probability

$$\mathrm{P}\{\mathbf{y} \in \mathcal{A}\} \triangleq \int_{\mathbb{R}^M} I_{\mathcal{A}}(\mathbf{y}')\, d\mu^{\mathbf{y}}(\mathbf{y}') = \int_{\mathcal{A}} d\mu^{\mathbf{y}}(\mathbf{y}'),$$

where $I_{\mathcal{A}}(\cdot) : \mathbb{R}^M \to \{0,1\}$ denotes the indicator function of the set $\mathcal{A}$. We will also consider a family of probability measures $\{\mu^{\mathbf{y}}_{\mathbf{x}}\}_{\mathbf{x}\in\mathcal{X}}$ parametrized by a nonrandom parameter vector $\mathbf{x} \in \mathcal{X}$. We assume that there exists a dominating measure $\mu_{\mathcal{E}}$, so that we can define the pdf $f(\mathbf{y};\mathbf{x})$ (again parametrized by $\mathbf{x}$) as the Radon-Nikodym derivative of the measure $\mu^{\mathbf{y}}_{\mathbf{x}}$ with respect to the measure $\mu_{\mathcal{E}}$ [1], [15]–[17]. (In general, we will choose for $\mu_{\mathcal{E}}$ the Lebesgue measure on $\mathbb{R}^M$.) We refer to both the set of measures $\{\mu^{\mathbf{y}}_{\mathbf{x}}\}_{\mathbf{x}\in\mathcal{X}}$ and the set of pdfs $\{f(\mathbf{y};\mathbf{x})\}_{\mathbf{x}\in\mathcal{X}}$ as the *statistical model*. Given a (possibly vector-valued) deterministic function $\mathbf{t}(\mathbf{y})$, the expectation operation is defined by [1]

$$\mathsf{E}_{\mathbf{x}}\{\mathbf{t}(\mathbf{y})\} \triangleq \int_{\mathbb{R}^M} \mathbf{t}(\mathbf{y}')\, d\mu^{\mathbf{y}}_{\mathbf{x}}(\mathbf{y}') = \int_{\mathbb{R}^M} \mathbf{t}(\mathbf{y}')\, f(\mathbf{y}';\mathbf{x})\, d\mathbf{y}',$$

where the subscript in $\mathsf{E}_{\mathbf{x}}$ indicates the dependence on the parameter vector $\mathbf{x}$ parametrizing $\mu^{\mathbf{y}}_{\mathbf{x}}(\mathbf{y})$ and $f(\mathbf{y};\mathbf{x})$.

II. Fundamentals

A. Review of MVE

It will be convenient to denote a classical (frequentist) estimation problem by the triple $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), \mathbf{g}(\cdot)\big)$, consisting of the parameter set $\mathcal{X} \subseteq \mathbb{R}^N$, the statistical model $\{f(\mathbf{y};\mathbf{x})\}_{\mathbf{x}\in\mathcal{X}}$, with $\mathbf{y} \in \mathbb{R}^M$, and the parameter function $\mathbf{g}(\cdot) : \mathcal{X} \to \mathbb{R}^P$. Note that our setting includes estimation of the parameter vector $\mathbf{x}$ itself, which is obtained when $\mathbf{g}(\mathbf{x}) = \mathbf{x}$. The result of estimating $\mathbf{g}(\mathbf{x})$ from $\mathbf{y}$ is an *estimate* $\hat{\mathbf{g}} \in \mathbb{R}^P$, which is derived from $\mathbf{y}$ via a deterministic *estimator* $\hat{\mathbf{g}}(\cdot)\colon \mathbb{R}^M \to \mathbb{R}^P$, i.e., $\hat{\mathbf{g}} = \hat{\mathbf{g}}(\mathbf{y})$. We assume that any estimator is a measurable mapping from $\mathbb{R}^M$ to $\mathbb{R}^P$ [1, Sec. 13]. A convenient characterization of the performance of an estimator $\hat{\mathbf{g}}(\cdot)$ is the *mean squared error* (MSE) defined as

$$\varepsilon \triangleq \mathsf{E}_{\mathbf{x}}\big\{\|\hat{\mathbf{g}}(\mathbf{y}) - \mathbf{g}(\mathbf{x})\|_2^2\big\} = \int_{\mathbb{R}^M} \|\hat{\mathbf{g}}(\mathbf{y}) - \mathbf{g}(\mathbf{x})\|_2^2\, f(\mathbf{y};\mathbf{x})\, d\mathbf{y}.$$

We will write $\varepsilon(\hat{\mathbf{g}}(\cdot);\mathbf{x})$ to explicitly indicate the dependence of the MSE on the estimator $\hat{\mathbf{g}}(\cdot)$ and the parameter vector $\mathbf{x}$. Unfortunately, for a general estimation problem $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), \mathbf{g}(\cdot)\big)$, there does not exist an estimator $\hat{\mathbf{g}}(\cdot)$ that minimizes the MSE simultaneously for all parameter vectors $\mathbf{x} \in \mathcal{X}$ [18], [19]. This follows from the fact that minimizing the MSE at a given parameter vector $\mathbf{x}_0$ always yields zero MSE; this is achieved by the estimator $\hat{\mathbf{g}}_0(\mathbf{y}) = \mathbf{g}(\mathbf{x}_0)$, which completely ignores the observation $\mathbf{y}$.

A popular rationale for the design of good estimators is MVE. This approach is based on the MSE decomposition

$$\varepsilon(\hat{\mathbf{g}}(\cdot);\mathbf{x}) = \|\mathbf{b}(\hat{\mathbf{g}}(\cdot);\mathbf{x})\|_2^2 + v(\hat{\mathbf{g}}(\cdot);\mathbf{x}), \tag{1}$$

with the *estimator bias* $\mathbf{b}(\hat{\mathbf{g}}(\cdot);\mathbf{x}) \triangleq \mathsf{E}_{\mathbf{x}}\{\hat{\mathbf{g}}(\mathbf{y})\} - \mathbf{g}(\mathbf{x})$ and the *estimator variance* $v(\hat{\mathbf{g}}(\cdot);\mathbf{x}) \triangleq \mathsf{E}_{\mathbf{x}}\{\|\hat{\mathbf{g}}(\mathbf{y}) - \mathsf{E}_{\mathbf{x}}\{\hat{\mathbf{g}}(\mathbf{y})\}\|_2^2\}$. In MVE, one fixes the bias for all parameter vectors, i.e., $\mathbf{b}(\hat{\mathbf{g}}(\cdot);\mathbf{x}) \overset{!}{=} \mathbf{c}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$, with a *prescribed bias function* $\mathbf{c}(\cdot) : \mathcal{X} \to \mathbb{R}^P$, and considers only estimators with the given bias. Note that fixing the estimator bias is equivalent to fixing the estimator mean, i.e., $\mathsf{E}_{\mathbf{x}}\{\hat{\mathbf{g}}(\mathbf{y})\} \overset{!}{=} \boldsymbol{\gamma}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$, with the *prescribed mean function* $\boldsymbol{\gamma}(\mathbf{x}) \triangleq \mathbf{c}(\mathbf{x}) + \mathbf{g}(\mathbf{x})$. The important special case of *unbiased estimation* is obtained for $\mathbf{c}(\mathbf{x}) \equiv \mathbf{0}$ or equivalently $\boldsymbol{\gamma}(\mathbf{x}) \equiv \mathbf{g}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. Fixing the bias can be viewed as a kind of regularization of the set of considered estimators [15], [19], because useless estimators like the estimator $\hat{\mathbf{g}}_0(\mathbf{y}) = \mathbf{g}(\mathbf{x}_0)$ are excluded. Another justification for fixing the bias is the fact that, if a large number of independent and identically distributed (i.i.d.) realizations $\{\mathbf{y}_i\}_{i=1}^L$ of the vector $\mathbf{y}$ are observed, then, under certain technical conditions, the bias term dominates in the decomposition (1). Thus, in that case, the MSE is small if and only if the bias is small; this means that the estimator has to be effectively unbiased, i.e., $\mathbf{b}(\hat{\mathbf{g}}(\cdot);\mathbf{x}) \approx \mathbf{0}$ for all $\mathbf{x} \in \mathcal{X}$.

For a fixed "reference" parameter vector $\mathbf{x}_0 \in \mathcal{X}$ and a prescribed bias function $\mathbf{c}(\cdot)$, we define the *set of allowed estimators* by

$$\mathcal{A}(\mathbf{c}(\cdot),\mathbf{x}_0) \triangleq \big\{\hat{\mathbf{g}}(\cdot) \,\big|\, v(\hat{\mathbf{g}}(\cdot);\mathbf{x}_0) < \infty\,,\, \mathbf{b}(\hat{\mathbf{g}}(\cdot);\mathbf{x}) = \mathbf{c}(\mathbf{x}) \;\forall \mathbf{x} \in \mathcal{X}\big\}\,.$$

We call a bias function $\mathbf{c}(\cdot)$ *valid* for the estimation problem $\mathcal{E} = (\mathcal{X}, f(\mathbf{y};\mathbf{x}), \mathbf{g}(\cdot))$ at $\mathbf{x}_0 \in \mathcal{X}$ if the set $\mathcal{A}(\mathbf{c}(\cdot),\mathbf{x}_0)$ is nonempty. This means that there is at least one estimator $\hat{\mathbf{g}}(\cdot)$ with finite variance at $\mathbf{x}_0$ and whose bias equals $\mathbf{c}(\cdot)$, i.e., $\mathbf{b}(\hat{\mathbf{g}}(\cdot);\mathbf{x}) = \mathbf{c}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. From (1), it follows that for a fixed bias $\mathbf{c}(\cdot)$, minimizing the MSE $\varepsilon(\hat{\mathbf{g}}(\cdot);\mathbf{x}_0)$ is equivalent to minimizing the variance $v(\hat{\mathbf{g}}(\cdot);\mathbf{x}_0)$. Therefore, in MVE, one attempts to find estimators that minimize the variance under the constraint of a prescribed bias function $\mathbf{c}(\cdot)$. Let

$$M(\mathbf{c}(\cdot),\mathbf{x}_0) \triangleq \inf_{\hat{\mathbf{g}}(\cdot) \in \mathcal{A}(\mathbf{c}(\cdot),\mathbf{x}_0)} v(\hat{\mathbf{g}}(\cdot);\mathbf{x}_0) \tag{2}$$

denote the minimum (strictly speaking, infimum) variance at $\mathbf{x}_0$ for bias function $\mathbf{c}(\cdot)$. If $\mathcal{A}(\mathbf{c}(\cdot),\mathbf{x}_0)$ is empty, i.e., if $\mathbf{c}(\cdot)$ is not valid, we set $M(\mathbf{c}(\cdot),\mathbf{x}_0) \triangleq \infty$. Any estimator $\hat{\mathbf{g}}^{(\mathbf{x}_0)}(\cdot) \in \mathcal{A}(\mathbf{c}(\cdot),\mathbf{x}_0)$ that achieves the infimum in (2), i.e., for which $v\big(\hat{\mathbf{g}}^{(\mathbf{x}_0)}(\cdot);\mathbf{x}_0\big) = M(\mathbf{c}(\cdot),\mathbf{x}_0)$, is called a *locally minimum variance* (LMV) estimator at $\mathbf{x}_0$ for bias function $\mathbf{c}(\cdot)$ [2], [3], [15]. The corresponding minimum variance $M(\mathbf{c}(\cdot),\mathbf{x}_0)$ is called the *minimum achievable variance* at $\mathbf{x}_0$ for bias function $\mathbf{c}(\cdot)$. The minimization problem (2) is referred to as a *minimum variance problem* (MVP). By its definition in (2), $M(\mathbf{c}(\cdot),\mathbf{x}_0)$ is a lower bound on the variance at $\mathbf{x}_0$ of any estimator with bias function $\mathbf{c}(\cdot)$, i.e.,

$$\hat{\mathbf{g}}(\cdot) \in \mathcal{A}(\mathbf{c}(\cdot),\mathbf{x}_0) \;\;\Rightarrow\;\; v(\hat{\mathbf{g}}(\cdot);\mathbf{x}_0) \geq M(\mathbf{c}(\cdot),\mathbf{x}_0)\,. \tag{3}$$

In fact, $M(\mathbf{c}(\cdot),\mathbf{x}_0)$ is the tightest lower bound, which is sometimes referred to as the *Barankin bound*.

If, for a prescribed bias function $\mathbf{c}(\cdot)$, there exists an estimator that is the LMV estimator *simultaneously* at all $\mathbf{x}_0 \in \mathcal{X}$, then that estimator is called the *uniformly minimum variance* (UMV) estimator for bias function $\mathbf{c}(\cdot)$ [2], [3], [15]. For many estimation problems, a UMV estimator does not exist. However, it always exists if there exists a *complete sufficient statistic* [15, Theorem 1.11 and Corollary 1.12], [20, Theorem 6.2.25]. Under mild conditions, this includes the case where the statistical model corresponds to an exponential family.

The variance to be minimized can be decomposed as

$$v(\hat{\mathbf{g}}(\cdot);\mathbf{x}_0) = \sum_{l\in[P]} v(\hat{g}_l(\cdot);\mathbf{x}_0)\,,$$

where $\hat{g}_l(\cdot) \triangleq \big(\hat{\mathbf{g}}(\cdot)\big)_l$ and $v(\hat{g}_l(\cdot);\mathbf{x}_0) \triangleq \mathsf{E}_{\mathbf{x}_0}\big\{\big[\hat{g}_l(\mathbf{y}) - \mathsf{E}_{\mathbf{x}_0}\{\hat{g}_l(\mathbf{y})\}\big]^2\big\}$ for $l \in [P]$. Moreover, $\hat{\mathbf{g}}(\cdot) \in \mathcal{A}(\mathbf{c}(\cdot),\mathbf{x}_0)$ if and only if $\hat{g}_l(\cdot) \in \mathcal{A}(c_l(\cdot),\mathbf{x}_0)$ for all $l \in [P]$, where $c_l(\cdot) \triangleq \big(\mathbf{c}(\cdot)\big)_l$. It follows that the minimization of $v(\hat{\mathbf{g}}(\cdot);\mathbf{x}_0)$ in (2) can be reduced to P separate problems of minimizing the component variances $v(\hat{g}_l(\cdot);\mathbf{x}_0)$, each involving the optimization of a single scalar component $\hat{g}_l(\cdot)$ of $\hat{\mathbf{g}}(\cdot)$ subject to the scalar bias constraint $b(\hat{g}_l(\cdot);\mathbf{x}) = c_l(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. Therefore, without loss of generality, we will hereafter assume that the parameter function $\mathbf{g}(\mathbf{x})$ is scalar-valued, i.e., $P = 1$.

B. Review of the RKHS Approach to MVE

A powerful mathematical toolbox for MVE is provided by RKHS theory [2], [3], [21]. In this subsection, we review basic definitions and results of RKHS theory and its application to MVE, and we discuss a differentiability property that will be relevant to the variance bounds considered in Section III.

An RKHS is associated with a *kernel function*, which is a function $R(\cdot\,,\cdot) \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the following two properties [21]:

- It is symmetric, i.e., $R(\mathbf{x}_1,\mathbf{x}_2) = R(\mathbf{x}_2,\mathbf{x}_1)$ for all $\mathbf{x}_1,\mathbf{x}_2 \in \mathcal{X}$;
- for every finite set $\{\mathbf{x}_1,\ldots,\mathbf{x}_D\} \subseteq \mathcal{X}$, the matrix $\mathbf{R} \in \mathbb{R}^{D\times D}$ with entries $R_{m,n} = R(\mathbf{x}_m,\mathbf{x}_n)$ is positive semidefinite.

There exists an RKHS for any kernel function $R(\cdot,\cdot) \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ [21]. This RKHS, denoted $\mathcal{H}(R)$, is a Hilbert space equipped with an inner product $\langle\cdot\,,\cdot\rangle_{\mathcal{H}(R)}$ such that, for any $\mathbf{x} \in \mathcal{X}$,

- $R(\cdot,\mathbf{x}) \in \mathcal{H}(R)$ (here, $R(\cdot,\mathbf{x})$ denotes the function $f_{\mathbf{x}}(\mathbf{x}') = R(\mathbf{x}',\mathbf{x})$ with a fixed $\mathbf{x} \in \mathcal{X}$);
- for any function $f(\cdot) \in \mathcal{H}(R)$,

$$\big\langle f(\cdot), R(\cdot,\mathbf{x})\big\rangle_{\mathcal{H}(R)} = f(\mathbf{x})\,. \tag{4}$$

Relation (4), which is known as the *reproducing property*, defines the inner product $\langle f,g\rangle_{\mathcal{H}(R)}$ for all $f(\cdot), g(\cdot) \in \mathcal{H}(R)$ because (in a certain sense) any $f(\cdot) \in \mathcal{H}(R)$ can be expanded into the set of functions $\{R(\cdot,\mathbf{x})\}_{\mathbf{x}\in\mathcal{X}}$. In particular, consider two functions $f(\cdot), g(\cdot) \in \mathcal{H}(R)$ that are given as $f(\cdot) = \sum_{\mathbf{x}_k\in\mathcal{D}} a_k R(\cdot,\mathbf{x}_k)$ and $g(\cdot) = \sum_{\mathbf{x}'_l\in\mathcal{D}'} b_l R(\cdot,\mathbf{x}'_l)$ with coefficients $a_k, b_l \in \mathbb{R}$ and (possibly infinite) sets $\mathcal{D}, \mathcal{D}' \subseteq \mathcal{X}$. Then, by the linearity of inner products and (4),

$$\big\langle f(\cdot), g(\cdot)\big\rangle_{\mathcal{H}(R)} = \sum_{\mathbf{x}_k\in\mathcal{D}} \sum_{\mathbf{x}'_l\in\mathcal{D}'} a_k b_l R(\mathbf{x}_k,\mathbf{x}'_l)\,.$$

1) The RKHS Associated with an MVP: Consider the class of MVPs that is defined by an estimation problem $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$, a reference parameter vector $\mathbf{x}_0 \in \mathcal{X}$, and all possible prescribed bias functions $c(\cdot) : \mathcal{X} \to \mathbb{R}$. With this class of MVPs, we can associate a kernel function $R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and, in turn, an RKHS $\mathcal{H}(R_{\mathcal{E},\mathbf{x}_0})$ [2], [3]. (Note that, as our notation indicates, $R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\cdot)$ and $\mathcal{H}(R_{\mathcal{E},\mathbf{x}_0})$ depend on $\mathcal{E}$ and $\mathbf{x}_0$ but not on $g(\cdot)$ or $c(\cdot)$.) We assume that

$$\mathrm{P}\{f(\mathbf{y};\mathbf{x}_0) \neq 0\} = 1\,, \tag{5}$$

where the probability is evaluated for some underlying dominating measure $\mu_{\mathcal{E}}$. We can then define the *likelihood ratio* as

$$\rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x}) \triangleq \begin{cases} \dfrac{f(\mathbf{y};\mathbf{x})}{f(\mathbf{y};\mathbf{x}_0)}\,, & \text{if } f(\mathbf{y};\mathbf{x}_0) \neq 0 \\ 0\,, & \text{else.} \end{cases} \tag{6}$$

We consider $\rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x})$ as a random variable (since it is a function of the random vector $\mathbf{y}$) that is parametrized by $\mathbf{x} \in \mathcal{X}$. Furthermore, we define the Hilbert space $\mathcal{L}_{\mathcal{E},\mathbf{x}_0}$ as the closure of the linear span[1] of the set of random variables $\big\{\rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x})\big\}_{\mathbf{x}\in\mathcal{X}}$. The topology of $\mathcal{L}_{\mathcal{E},\mathbf{x}_0}$ is determined by the inner product $\langle\cdot\,,\cdot\rangle_{\mathrm{RV}} : \mathcal{L}_{\mathcal{E},\mathbf{x}_0} \times \mathcal{L}_{\mathcal{E},\mathbf{x}_0} \to \mathbb{R}$ defined by

$$\begin{aligned}\big\langle \rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x}_1), \rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x}_2)\big\rangle_{\mathrm{RV}}& \\ \triangleq \mathsf{E}_{\mathbf{x}_0}\big\{\rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x}_1)\, \rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x}_2)\big\}&\,. \end{aligned}\tag{7}$$

It can be shown that it is sufficient to define the inner product only for the random variables $\big\{\rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x})\big\}_{\mathbf{x}\in\mathcal{X}}$ [2]. We will assume that

$$\big\langle \rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x}_1), \rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x}_2)\big\rangle_{\mathrm{RV}} < \infty\,, \quad \text{for all } \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}. \tag{8}$$

The assumptions (5) and (8) (or variants thereof) are standard in the literature on MVE [2], [3], [23], [24]. They are typically satisfied for the important and large class of estimation problems arising from exponential families (cf. Section V).

The inner product $\langle\cdot\,,\cdot\rangle_{\mathrm{RV}} : \mathcal{L}_{\mathcal{E},\mathbf{x}_0} \times \mathcal{L}_{\mathcal{E},\mathbf{x}_0} \to \mathbb{R}$ can now be interpreted as a kernel function $R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$:

$$R_{\mathcal{E},\mathbf{x}_0}(\mathbf{x}_1,\mathbf{x}_2) \triangleq \big\langle \rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x}_1), \rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x}_2)\big\rangle_{\mathrm{RV}}\,. \tag{9}$$

The RKHS induced by $R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\cdot)$ will be denoted by $\mathcal{H}_{\mathcal{E},\mathbf{x}_0}$, i.e., $\mathcal{H}_{\mathcal{E},\mathbf{x}_0} \triangleq \mathcal{H}(R_{\mathcal{E},\mathbf{x}_0})$. This is the RKHS associated with the estimation problem $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ and the corresponding class of MVPs at $\mathbf{x}_0 \in \mathcal{X}$. (Note that $R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\cdot)$ and $\mathcal{H}_{\mathcal{E},\mathbf{x}_0}$ do not depend on $g(\cdot)$.)

Assumption (5) implies that the likelihood ratio $\rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x})$ is measurable with respect to the underlying dominating measure $\mu_{\mathcal{E}}$. Furthermore, $\rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x})$ is the Radon-Nikodym derivative [1], [16] of the probability measure $\mu^{\mathbf{y}}_{\mathbf{x}}$ induced by $f(\mathbf{y};\mathbf{x})$ with respect to the probability measure $\mu^{\mathbf{y}}_{\mathbf{x}_0}$ induced by $f(\mathbf{y};\mathbf{x}_0)$ (cf. [1], [22], [25]). It is important to observe that $\rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x})$ does not depend on the dominating measure $\mu_{\mathcal{E}}$ underlying the definition of the pdfs $f(\mathbf{y};\mathbf{x})$. Thus, the kernel $R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\cdot)$ given by (9) does not depend on $\mu_{\mathcal{E}}$ either. Moreover, under assumption (5), there exists the Radon-Nikodym derivative of $\mu^{\mathbf{y}}_{\mathbf{x}}$ with respect to $\mu^{\mathbf{y}}_{\mathbf{x}_0}$. This implies, trivially, that $\mu^{\mathbf{y}}_{\mathbf{x}}$ is absolutely continuous with respect to $\mu^{\mathbf{y}}_{\mathbf{x}_0}$, or, equivalently, that $\mu^{\mathbf{y}}_{\mathbf{x}_0}$ dominates the measures $\{\mu^{\mathbf{y}}_{\mathbf{x}}\}_{\mathbf{x}\in\mathcal{X}}$ [1, p. 443]. Therefore, we can use the measure $\mu^{\mathbf{y}}_{\mathbf{x}_0}$ as the base measure $\mu_{\mathcal{E}}$ for the estimation problem $\mathcal{E}$.

The two Hilbert spaces $\mathcal{L}_{\mathcal{E},\mathbf{x}_0}$ and $\mathcal{H}_{\mathcal{E},\mathbf{x}_0}$ are isometric. In fact, as proven in [2], a specific congruence (i.e., isometric mapping of functions in $\mathcal{H}_{\mathcal{E},\mathbf{x}_0}$ to functions in $\mathcal{L}_{\mathcal{E},\mathbf{x}_0}$) $\mathsf{J}[\cdot] : \mathcal{H}_{\mathcal{E},\mathbf{x}_0} \to \mathcal{L}_{\mathcal{E},\mathbf{x}_0}$ is given by

$$\mathsf{J}[R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\mathbf{x})] = \rho_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\mathbf{x})\,.$$

The isometry $\mathsf{J}[f(\cdot)]$ can be evaluated for an arbitrary function $f(\cdot) \in \mathcal{H}_{\mathcal{E},\mathbf{x}_0}$ by expanding $f(\cdot)$ into the elementary functions $\{R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\mathbf{x})\}_{\mathbf{x}\in\mathcal{X}}$ (cf. [2]). Given the expansion $f(\cdot) = \sum_{\mathbf{x}_k\in\mathcal{D}} a_k R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\mathbf{x}_k)$ with coefficients $a_k \in \mathbb{R}$ and a (possibly infinite) set $\mathcal{D} \subseteq \mathcal{X}$, the isometric image of $f(\cdot)$ is obtained as $\mathsf{J}[f(\cdot)] = \sum_{\mathbf{x}_k\in\mathcal{D}} a_k\, \rho_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\mathbf{x}_k)$.

2) RKHS-based Analysis of MVE: An RKHS-based analysis of MVE is enabled by the following central result. Consider an estimation problem $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$, a fixed reference parameter vector $\mathbf{x}_0 \in \mathcal{X}$, and a prescribed bias function $c(\cdot) : \mathcal{X} \to \mathbb{R}$, corresponding to the prescribed mean function $\gamma(\cdot) \triangleq c(\cdot) + g(\cdot)$. Then, as shown in [2] and [3], the following holds:

- The bias function $c(\cdot)$ is valid for $\mathcal{E}$ at $\mathbf{x}_0$ if and only if $\gamma(\cdot)$ belongs to the RKHS $\mathcal{H}_{\mathcal{E},\mathbf{x}_0}$, i.e.,

$$\mathcal{A}(c(\cdot),\mathbf{x}_0) \neq \emptyset \iff \gamma(\cdot) \in \mathcal{H}_{\mathcal{E},\mathbf{x}_0}\,. \tag{10}$$

- If the bias function $c(\cdot)$ is valid, the corresponding minimum achievable variance at $\mathbf{x}_0$ is given by

$$M(c(\cdot),\mathbf{x}_0) \,=\, \|\gamma(\cdot)\|^2_{\mathcal{H}_{\mathcal{E},\mathbf{x}_0}} - \gamma^2(\mathbf{x}_0)\,, \tag{11}$$

  and the LMV estimator at $\mathbf{x}_0$ is given by

$$\hat{g}^{(\mathbf{x}_0)}(\cdot) \,=\, \mathsf{J}[\gamma(\cdot)]\,. \tag{12}$$

This result shows that the RKHS $\mathcal{H}_{\mathcal{E},\mathbf{x}_0}$ is equal to the set of the mean functions $\gamma(\mathbf{x}) = \mathsf{E}_{\mathbf{x}}\{\hat{g}(\mathbf{y})\}$ of all estimators $\hat{g}(\cdot)$ with a finite variance at $\mathbf{x}_0$, i.e., $v(\hat{g}(\cdot);\mathbf{x}_0) < \infty$. Furthermore, the problem of solving the MVP (2) can be reduced to the computation of the squared norm $\|\gamma(\cdot)\|^2_{\mathcal{H}_{\mathcal{E},\mathbf{x}_0}}$ and the isometric image $\mathsf{J}[\gamma(\cdot)]$ of the prescribed mean function $\gamma(\cdot)$, viewed as an element of the RKHS $\mathcal{H}_{\mathcal{E},\mathbf{x}_0}$. This is especially helpful if a simple characterization of $\mathcal{H}_{\mathcal{E},\mathbf{x}_0}$ is available. Here, following the terminology of [3], what is meant by "simple characterization" is the availability of an orthonormal basis (ONB) for $\mathcal{H}_{\mathcal{E},\mathbf{x}_0}$ such that the inner products of $\gamma(\cdot)$ with the ONB functions can be computed easily.

If such an ONB of $\mathcal{H}_{\mathcal{E},\mathbf{x}_0}$ cannot be found, the relation (11) can still be used to derive lower bounds on the minimum achievable variance $M(c(\cdot),\mathbf{x}_0)$. Indeed, because of (11), any lower bound on $\|\gamma(\cdot)\|^2_{\mathcal{H}_{\mathcal{E},\mathbf{x}_0}}$ induces a lower bound on $M(c(\cdot),\mathbf{x}_0)$. A large class of lower bounds on $\|\gamma(\cdot)\|^2_{\mathcal{H}_{\mathcal{E},\mathbf{x}_0}}$ can be obtained via projections of $\gamma(\cdot)$ onto a subspace $\mathcal{U} \subseteq \mathcal{H}_{\mathcal{E},\mathbf{x}_0}$. Denoting the orthogonal projection of $\gamma(\cdot)$ onto $\mathcal{U}$ by $\gamma_{\mathcal{U}}(\cdot)$,

[1]A detailed discussion of the concepts of closure, inner product, orthonormal basis, and linear span in the context of abstract Hilbert space theory can be found in [2] and [22].

we have $\|\gamma_{\mathcal{U}}(\cdot)\|^2_{\mathcal{H}_{\mathcal{E},\mathbf{x}_0}} \leq \|\gamma(\cdot)\|^2_{\mathcal{H}_{\mathcal{E},\mathbf{x}_0}}$ [22, Chapter 4] and thus, from (11),

$$M(c(\cdot),\mathbf{x}_0) \,\geq\, \|\gamma_{\mathcal{U}}(\cdot)\|^2_{\mathcal{H}_{\mathcal{E},\mathbf{x}_0}} - \gamma^2(\mathbf{x}_0)\,, \tag{13}$$

for an arbitrary subspace $\mathcal{U} \subseteq \mathcal{H}_{\mathcal{E},\mathbf{x}_0}$. In particular, let us consider the special case of a finite-dimensional subspace $\mathcal{U} \subseteq \mathcal{H}_{\mathcal{E},\mathbf{x}_0}$ that is spanned by a given set of functions $u_l(\cdot) \in \mathcal{H}_{\mathcal{E},\mathbf{x}_0}$, i.e.,

$$\mathcal{U} \,=\, \mathrm{span}\{u_l(\cdot)\}_{l\in[L]} \,\triangleq\, \left\{ f(\cdot) = \sum_{l\in[L]} a_l u_l(\cdot) \,\middle|\, a_l \in \mathbb{R} \right\}. \tag{14}$$

Here, $\|\gamma_{\mathcal{U}}(\cdot)\|^2_{\mathcal{H}_{\mathcal{E},\mathbf{x}_0}}$ can be evaluated very easily due to the following expression [10, Theorem 3.1.8]:

$$\|\gamma_{\mathcal{U}}(\cdot)\|^2_{\mathcal{H}_{\mathcal{E},\mathbf{x}_0}} = \boldsymbol{\gamma}^T \mathbf{G}^{\dagger} \boldsymbol{\gamma}\,, \tag{15}$$

where the vector $\boldsymbol{\gamma} \in \mathbb{R}^L$ and the matrix $\mathbf{G} \in \mathbb{R}^{L\times L}$ are given elementwise by

$$\gamma_l \,=\, \langle \gamma(\cdot), u_l(\cdot)\rangle_{\mathcal{H}_{\mathcal{E},\mathbf{x}_0}}\,, \quad G_{l,l'} \,=\, \langle u_l(\cdot), u_{l'}(\cdot)\rangle_{\mathcal{H}_{\mathcal{E},\mathbf{x}_0}}\,. \tag{16}$$

If all $u_l(\cdot)$ are linearly independent, then a larger number L of basis functions $u_l(\cdot)$ entails a higher dimension of $\mathcal{U}$ and, thus, a larger $\|\gamma_{\mathcal{U}}(\cdot)\|^2_{\mathcal{H}_{\mathcal{E},\mathbf{x}_0}}$; this implies that the lower bound (13) will be higher (i.e., tighter). In Section III, we will show that some well-known lower bounds on the estimator variance are obtained from (13) and (15), using a subspace $\mathcal{U}$ of the form (14) and specific choices for the functions $u_l(\cdot)$ spanning $\mathcal{U}$.

3) Regular Estimation Problems and Differentiable RKHS: Some of the lower bounds to be considered in Section III require the estimation problem to satisfy certain regularity conditions.

Definition II.1. *An estimation problem* $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ *satisfying* (8) *is said to be* regular up to order $m \in \mathbb{N}$ at an interior point $\mathbf{x}_0 \in \mathcal{X}^{\mathrm{o}}$ *if the following holds:*

- *For every multi-index* $\mathbf{p} \in \mathbb{Z}_+^N$ *with entries* $p_k \leq m$, *the partial derivatives* $\frac{\partial^{\mathbf{p}} f(\mathbf{y};\mathbf{x})}{\partial \mathbf{x}^{\mathbf{p}}}$ *exist and satisfy*

$$\mathsf{E}_{\mathbf{x}_0}\left\{ \left(\frac{1}{f(\mathbf{y};\mathbf{x}_0)} \frac{\partial^{\mathbf{p}} f(\mathbf{y};\mathbf{x})}{\partial \mathbf{x}^{\mathbf{p}}} \right)^2 \right\} < \infty\,, \quad \textit{for all } \mathbf{x} \in \mathcal{B}(\mathbf{x}_0, r)\,,$$

  where $r > 0$ *is a suitably chosen radius such that* $\mathcal{B}(\mathbf{x}_0, r) \subseteq \mathcal{X}$.
- *For any function* $h(\cdot) : \mathbb{R}^M \to \mathbb{R}$ *such that* $\mathsf{E}_{\mathbf{x}}\{h(\mathbf{y})\}$ *exists, the expectation operation commutes with partial differentiation in the sense that, for every multi-index* $\mathbf{p} \in \mathbb{Z}_+^N$ *with* $p_k \leq m$,

$$\frac{\partial^{\mathbf{p}}}{\partial \mathbf{x}^{\mathbf{p}}} \int_{\mathbb{R}^M} h(\mathbf{y})\, f(\mathbf{y};\mathbf{x})\, d\mathbf{y} = \int_{\mathbb{R}^M} h(\mathbf{y})\, \frac{\partial^{\mathbf{p}} f(\mathbf{y};\mathbf{x})}{\partial \mathbf{x}^{\mathbf{p}}}\, d\mathbf{y}\,, \quad \textit{for all } \mathbf{x} \in \mathcal{B}(\mathbf{x}_0, r)\,, \tag{17}$$

  or equivalently

$$\frac{\partial^{\mathbf{p}} \mathsf{E}_{\mathbf{x}}\{h(\mathbf{y})\}}{\partial \mathbf{x}^{\mathbf{p}}} \,=\, \mathsf{E}_{\mathbf{x}}\left\{ h(\mathbf{y}) \frac{1}{f(\mathbf{y};\mathbf{x})} \frac{\partial^{\mathbf{p}} f(\mathbf{y};\mathbf{x})}{\partial \mathbf{x}^{\mathbf{p}}} \right\}, \quad \textit{for all } \mathbf{x} \in \mathcal{B}(\mathbf{x}_0, r)\,, \tag{18}$$

  provided that the right hand side of (17) *and* (18) *is finite.*
- *For every pair of multi-indices* $\mathbf{p}_1, \mathbf{p}_2 \in \mathbb{Z}_+^N$ *with* $p_{1,k} \leq m$ *and* $p_{2,k} \leq m$, *the expectation*

$$\mathsf{E}_{\mathbf{x}_0}\left\{ \frac{1}{f^2(\mathbf{y};\mathbf{x}_0)} \frac{\partial^{\mathbf{p}_1} f(\mathbf{y};\mathbf{x}_1)}{\partial \mathbf{x}_1^{\mathbf{p}_1}} \frac{\partial^{\mathbf{p}_2} f(\mathbf{y};\mathbf{x}_2)}{\partial \mathbf{x}_2^{\mathbf{p}_2}} \right\}$$

  depends continuously on the parameter vectors $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{B}(\mathbf{x}_0, r)$.

We remark that the notion of a *regular estimation problem* according to Definition II.1 is somewhat similar to the notion of a *regular statistical experiment* introduced in [17, Section I.7].

The RKHS $\mathcal{H}_{\mathcal{E},\mathbf{x}_0}$ associated with a regular estimation problem $\mathcal{E}$ has an important structural property, which we will term *differentiable*. More precisely, we call an RKHS $\mathcal{H}(R)$ *differentiable up to order* m if it is associated with a kernel $R(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that is differentiable up to a given order m. The properties of differentiable RKHSs have been previously studied, e.g., in [26]–[28]. As shown in [10, Theorem 4.4.3], if $\mathcal{E}$ is regular up to order m at $\mathbf{x}_0$, then $\mathcal{H}_{\mathcal{E},\mathbf{x}_0}$ is differentiable up to order m.

It will be seen that, under certain conditions, the functions belonging to an RKHS $\mathcal{H}(R)$ that is differentiable up to any order are characterized completely by their partial derivatives at any point $\mathbf{x}_0 \in \mathcal{X}^{\mathrm{o}}$. This implies via (10) together with identity (20) below that, for a regular estimation problem, the mean function $\gamma(\mathbf{x}) = \mathsf{E}_{\mathbf{x}}\{\hat{g}(\mathbf{y})\}$ of any estimator $\hat{g}(\cdot)$ with finite variance at $\mathbf{x}_0$ is completely specified by the partial derivatives $\big\{\frac{\partial^{\mathbf{p}}\gamma(\mathbf{x})}{\partial \mathbf{x}^{\mathbf{p}}}\big|_{\mathbf{x}=\mathbf{x}_0}\big\}_{\mathbf{p}\in\mathbb{Z}_+^N}$ (cf. Lemma V.3 in Section V-D).

Further important properties of a differentiable RKHS have been reported in [9] and [27]. In particular, for an RKHS $\mathcal{H}(R)$ that is differentiable up to order m, and for any $\mathbf{x}_0 \in \mathcal{X}^{\mathrm{o}}$ and any $\mathbf{p} \in \mathbb{Z}_+^N$ with $p_k \leq m$, the following holds:

- The function $r^{(\mathbf{p})}_{\mathbf{x}_0}(\cdot) \colon \mathcal{X} \to \mathbb{R}$ defined by

$$r^{(\mathbf{p})}_{\mathbf{x}_0}(\mathbf{x}) \,\triangleq\, \frac{\partial^{\mathbf{p}} R(\mathbf{x},\mathbf{x}_2)}{\partial \mathbf{x}_2^{\mathbf{p}}}\bigg|_{\mathbf{x}_2=\mathbf{x}_0} \tag{19}$$

  is an element of $\mathcal{H}(R)$, i.e., $r^{(\mathbf{p})}_{\mathbf{x}_0}(\cdot) \in \mathcal{H}(R)$.
- For any function $f(\cdot) \in \mathcal{H}(R)$, the partial derivative $\frac{\partial^{\mathbf{p}} f(\mathbf{x})}{\partial \mathbf{x}^{\mathbf{p}}}\big|_{\mathbf{x}=\mathbf{x}_0}$ exists.
- The inner product of $r^{(\mathbf{p})}_{\mathbf{x}_0}(\cdot)$ with an arbitrary function $f(\cdot) \in \mathcal{H}(R)$ is given by

$$\big\langle r^{(\mathbf{p})}_{\mathbf{x}_0}(\cdot), f(\cdot) \big\rangle_{\mathcal{H}(R)} \,=\, \frac{\partial^{\mathbf{p}} f(\mathbf{x})}{\partial \mathbf{x}^{\mathbf{p}}}\bigg|_{\mathbf{x}=\mathbf{x}_0}. \tag{20}$$

Thus, an RKHS $\mathcal{H}(R)$ that is differentiable up to order m contains the functions $\big\{r^{(\mathbf{p})}_{\mathbf{x}_0}(\mathbf{x})\big\}_{p_k\leq m}$, and the inner products of any function $f(\cdot) \in \mathcal{H}(R)$ with the $r^{(\mathbf{p})}_{\mathbf{x}_0}(\mathbf{x})$ can be computed easily via differentiation of $f(\cdot)$. This makes function sets $\big\{r^{(\mathbf{p})}_{\mathbf{x}_0}(\mathbf{x})\big\}$ appear as interesting candidates for a simple characterization of the RKHS $\mathcal{H}(R)$. However, in general, these function sets are not guaranteed to be complete or orthonormal, i.e., they do not generally constitute an ONB. An important exception is given by certain estimation problems $\mathcal{E}$ involving an exponential family of distributions, which will be studied in Section V-D.

Consider an estimation problem $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ that is regular up to order $m \in \mathbb{N}$ at $\mathbf{x}_0 \in \mathcal{X}^{\mathrm{o}}$. According to (10), the mean function $\gamma(\cdot)$ of any estimator with finite variance at $\mathbf{x}_0$ belongs to the RKHS $\mathcal{H}_{\mathcal{E},\mathbf{x}_0}$. Since $\mathcal{E}$ is assumed regular up to order m, $\mathcal{H}_{\mathcal{E},\mathbf{x}_0}$ is differentiable up to order m. This, in turn, implies[2] via (10) and (20) that the partial derivatives of $\gamma(\cdot)$ at $\mathbf{x}_0$ exist up to order m. Therefore, for the derivation of lower bounds on the minimum achievable variance at $\mathbf{x}_0$ in the case of an estimation problem that is regular up to order m at $\mathbf{x}_0$, we can always tacitly assume that the partial derivatives of $\gamma(\cdot)$ at $\mathbf{x}_0$ exist up to order m; otherwise the corresponding bias function $c(\cdot) = \gamma(\cdot) - g(\cdot)$ would not be valid, i.e., there would not exist any estimator with mean function $\gamma(\cdot)$ (or, equivalently, bias function $c(\cdot)$) and finite variance at $\mathbf{x}_0$.

III. RKHS Formulation of Known Variance Bounds

Consider an estimation problem $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ and an estimator $\hat{g}(\cdot)$ with mean function $\gamma(\mathbf{x}) = \mathsf{E}_{\mathbf{x}}\{\hat{g}(\mathbf{y})\}$ and bias function $c(\mathbf{x}) = \gamma(\mathbf{x}) - g(\mathbf{x})$. We assume that $\hat{g}(\cdot)$ has a finite variance at $\mathbf{x}_0$, which implies that $c(\cdot)$ is valid and $\hat{g}(\cdot) \in \mathcal{A}(c(\cdot), \mathbf{x}_0)$; hence, $\mathcal{A}(c(\cdot), \mathbf{x}_0)$ is nonempty. Then, $\gamma(\cdot) \in \mathcal{H}_{\mathcal{E},\mathbf{x}_0}$ according to (10). We also recall from our discussion further above that if the estimation problem $\mathcal{E}$ is regular at $\mathbf{x}_0$ up to order m, then the partial derivatives $\frac{\partial^{\mathbf{p}} \gamma(\mathbf{x})}{\partial \mathbf{x}^{\mathbf{p}}}\big|_{\mathbf{x}=\mathbf{x}_0}$ exist for all $\mathbf{p} \in \mathbb{Z}_+^N$ with $p_k \leq m$.

In this section, we will demonstrate how five known lower bounds on the variance—Barankin bound, Cramér–Rao bound, constrained Cramér–Rao bound, Bhattacharyya bound, and Hammersley-Chapman-Robbins bound—can be formulated in a unified manner within the RKHS framework. More specifically, by combining (3) with (13), it follows that the variance of $\hat{g}(\cdot)$ at $\mathbf{x}_0$ is lower bounded as

$$v(\hat{g}(\cdot);\mathbf{x}_0) \,\geq\, \|\gamma_{\mathcal{U}}(\cdot)\|^2_{\mathcal{H}_{\mathcal{E},\mathbf{x}_0}} - \gamma^2(\mathbf{x}_0)\,, \tag{21}$$

where $\mathcal{U}$ is any subspace of $\mathcal{H}_{\mathcal{E},\mathbf{x}_0}$. The five variance bounds to be considered are obtained via specific choices of $\mathcal{U}$.

A. Barankin Bound

For a (valid) prescribed bias function $c(\cdot)$, the Barankin bound [23], [29] is the minimum achievable variance at $\mathbf{x}_0$, i.e., the variance of the LMV estimator at $\mathbf{x}_0$, which we denoted $M(c(\cdot), \mathbf{x}_0)$. This is the tightest lower bound on the variance, cf. (3). Using the RKHS expression of $M(c(\cdot), \mathbf{x}_0)$ in (11), the Barankin bound can be written as

$$v(\hat{g}(\cdot);\mathbf{x}_0) \,\geq\, M(c(\cdot), \mathbf{x}_0) \,=\, \|\gamma(\cdot)\|^2_{\mathcal{H}_{\mathcal{E},\mathbf{x}_0}} - \gamma^2(\mathbf{x}_0)\,, \tag{22}$$

with $\gamma(\cdot) = c(\cdot) + g(\cdot)$, for any estimator $\hat{g}(\cdot)$ with bias function $c(\cdot)$. Comparing with (21), we see that the Barankin bound is obtained for the special choice $\mathcal{U} = \mathcal{H}_{\mathcal{E},\mathbf{x}_0}$, in which case $\gamma_{\mathcal{U}}(\cdot) = \gamma(\cdot)$ and (21) reduces to (22).

In the literature [23], [29], the following expression of the Barankin bound is usually considered. Let $\mathcal{D} \triangleq \{\mathbf{x}_1, \ldots, \mathbf{x}_L\} \subseteq \mathcal{X}$ with finite size $L = |\mathcal{D}| \in \mathbb{N}$, and let $\mathbf{a} \triangleq (a_1 \cdots a_L)^T$ with $a_l \in \mathbb{R}$. Then the Barankin bound can be written as [23, Theorem 4]

$$\begin{aligned} v(\hat{g}(\cdot);\mathbf{x}_0) &\\ \geq\, & M(c(\cdot), \mathbf{x}_0) \\ =\, & \sup_{\mathcal{D}\subseteq\mathcal{X},\, L\in\mathbb{N},\, \mathbf{a}\in\mathcal{A}_{\mathcal{D}}} \frac{\Big(\sum_{l\in[L]} a_l\, [\gamma(\mathbf{x}_l) - \gamma(\mathbf{x}_0)]\Big)^2}{\mathsf{E}_{\mathbf{x}_0}\Big\{\Big(\sum_{l\in[L]} a_l\, \rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x}_l)\Big)^2\Big\}}\,, \end{aligned} \tag{23}$$

where $\rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x}_l)$ is the likelihood ratio as defined in (6) and $\mathcal{A}_{\mathcal{D}}$ is defined as the set of all $\mathbf{a} \in \mathbb{R}^L$ for which the denominator $\mathsf{E}_{\mathbf{x}_0}\big\{\big(\sum_{l\in[L]} a_l\, \rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x}_l)\big)^2\big\}$ does not vanish. Note that our notation $\sup_{\mathcal{D}\subseteq\mathcal{X},\, L\in\mathbb{N},\, \mathbf{a}\in\mathcal{A}_{\mathcal{D}}}$ is intended to indicate that the supremum is taken not only with respect to the elements $\mathbf{x}_l$ of $\mathcal{D}$ but also the number of elements (size of $\mathcal{D}$), L. We will now verify that the bound in (23) can be obtained from our RKHS expression in (22). We will use the following result derived in [10, Theorem 3.1.2].

Lemma III.1. *Consider an RKHS $\mathcal{H}(R)$ with kernel $R(\cdot,\cdot)\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$. Let $\mathcal{D} \triangleq \{\mathbf{x}_1, \ldots, \mathbf{x}_L\} \subseteq \mathcal{X}$ with some $L = |\mathcal{D}| \in \mathbb{N}$, and let $\mathbf{a} \triangleq (a_1 \cdots a_L)^T$ with $a_l \in \mathbb{R}$. Then the norm $\|f(\cdot)\|_{\mathcal{H}(R)}$ of any function $f(\cdot) \in \mathcal{H}(R)$ can be expressed as*

$$\|f(\cdot)\|_{\mathcal{H}(R)} \,=\, \sup_{\mathcal{D}\subseteq\mathcal{X},\, L\in\mathbb{N},\, \mathbf{a}\in\mathcal{A}'_{\mathcal{D}}} \frac{\sum_{l\in[L]} a_l f(\mathbf{x}_l)}{\sqrt{\sum_{l,l'\in[L]} a_l a_{l'} R(\mathbf{x}_l, \mathbf{x}_{l'})}}\,, \tag{24}$$

where $\mathcal{A}'_{\mathcal{D}}$ is the set of all $\mathbf{a} \in \mathbb{R}^L$ for which $\sum_{l,l'\in[L]} a_l a_{l'} \times R(\mathbf{x}_l, \mathbf{x}_{l'})$ does not vanish.

We will furthermore use the fact—shown in [10, Section 2.3.5]—that $M(c(\cdot), \mathbf{x}_0)$ remains unchanged when the prescribed mean function $\gamma(\mathbf{x})$ is replaced by $\tilde{\gamma}(\mathbf{x}) \triangleq \gamma(\mathbf{x}) + c$ with an arbitrary constant c. Setting in particular $c = -\gamma(\mathbf{x}_0)$, we have $\tilde{\gamma}(\mathbf{x}) = \gamma(\mathbf{x}) - \gamma(\mathbf{x}_0)$ and $\tilde{\gamma}(\mathbf{x}_0) = 0$, and thus (22) simplifies to

$$v(\hat{g}(\cdot);\mathbf{x}_0) \,\geq\, M(c(\cdot), \mathbf{x}_0) \,=\, \|\tilde{\gamma}(\cdot)\|^2_{\mathcal{H}_{\mathcal{E},\mathbf{x}_0}}\,. \tag{25}$$

Using (24) in (25), we obtain

$$\begin{aligned} M(c(\cdot), \mathbf{x}_0) &= \sup_{\mathcal{D}\subseteq\mathcal{X},\, L\in\mathbb{N},\, \mathbf{a}\in\mathcal{A}'_{\mathcal{D}}} \frac{\Big(\sum_{l\in[L]} a_l\, \tilde{\gamma}(\mathbf{x}_l)\Big)^2}{\sum_{l,l'\in[L]} a_l a_{l'} R_{\mathcal{E},\mathbf{x}_0}(\mathbf{x}_l, \mathbf{x}_{l'})} \\ &= \sup_{\mathcal{D}\subseteq\mathcal{X},\, L\in\mathbb{N},\, \mathbf{a}\in\mathcal{A}'_{\mathcal{D}}} \frac{\Big(\sum_{l\in[L]} a_l\, [\gamma(\mathbf{x}_l) - \gamma(\mathbf{x}_0)]\Big)^2}{\sum_{l,l'\in[L]} a_l a_{l'} R_{\mathcal{E},\mathbf{x}_0}(\mathbf{x}_l, \mathbf{x}_{l'})}\,. \end{aligned} \tag{26}$$

From (9) and (7), we have $R_{\mathcal{E},\mathbf{x}_0}(\mathbf{x}_1, \mathbf{x}_2) = \mathsf{E}_{\mathbf{x}_0}\big\{\rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x}_1) \times \rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x}_2)\big\}$, and thus the denominator in (26) becomes

$$\begin{aligned} &\sum_{l,l'\in[L]} a_l a_{l'} R_{\mathcal{E},\mathbf{x}_0}(\mathbf{x}_l, \mathbf{x}_{l'}) \\ &\qquad = \mathsf{E}_{\mathbf{x}_0}\Bigg\{ \sum_{l,l'\in[L]} a_l a_{l'} \rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x}_l)\, \rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x}_{l'}) \Bigg\} \end{aligned}$$

[2] Indeed, it follows from (10) that the mean function $\gamma(\cdot)$ belongs to the RKHS $\mathcal{H}_{\mathcal{E},\mathbf{x}_0}$. Therefore, by (20), the partial derivatives of $\gamma(\cdot)$ at $\mathbf{x}_0$ coincide with well-defined inner products of functions in $\mathcal{H}_{\mathcal{E},\mathbf{x}_0}$.

$$= \mathsf{E}_{\mathbf{x}_0}\left\{\left(\sum_{l\in[L]} a_l\, \rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x}_l)\right)^2\right\},$$

whence it also follows that $\mathcal{A}'_{\mathcal{D}} = \mathcal{A}_{\mathcal{D}}$. Therefore, (26) is equivalent to (23). Hence, we have shown that our RKHS expression (22) is equivalent to (23).

B. Cramér–Rao Bound

The Cramér–Rao bound (CRB) [18], [30], [31] is the most popular lower variance bound. Since the CRB applies to any estimator with a prescribed bias function $c(\cdot)$, it yields also a lower bound on the minimum achievable variance $M(c(\cdot),\mathbf{x}_0)$ (cf. (3)).

Consider an estimation problem $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ that is regular up to order 1 at $\mathbf{x}_0 \in \mathcal{X}^{\mathrm{o}}$ in the sense of Definition II.1. Let $\hat{g}(\cdot)$ denote an estimator with mean function $\gamma(\mathbf{x}) = \mathsf{E}_{\mathbf{x}}\{\hat{g}(\mathbf{y})\}$ and finite variance at $\mathbf{x}_0$ (i.e., $v(\hat{g}(\cdot);\mathbf{x}_0) < \infty$). Then, this variance is lower bounded by the CRB

$$v(\hat{g}(\cdot);\mathbf{x}_0) \,\geq\, \mathbf{b}^T(\mathbf{x}_0)\,\mathbf{J}^{\dagger}(\mathbf{x}_0)\,\mathbf{b}(\mathbf{x}_0)\,, \tag{27}$$

where $\mathbf{b}(\mathbf{x}_0) \triangleq \frac{\partial\gamma(\mathbf{x})}{\partial\mathbf{x}}\big|_{\mathbf{x}=\mathbf{x}_0}$ and $\mathbf{J}(\mathbf{x}_0) \in \mathbb{R}^{N\times N}$, known as the *Fisher information matrix* associated with $\mathcal{E}$, is given elementwise by

$$\big(\mathbf{J}(\mathbf{x}_0)\big)_{k,l} \triangleq \mathsf{E}_{\mathbf{x}_0}\left\{\frac{\partial \log f(\mathbf{y};\mathbf{x})}{\partial x_k}\,\frac{\partial \log f(\mathbf{y};\mathbf{x})}{\partial x_l}\bigg|_{\mathbf{x}=\mathbf{x}_0}\right\}. \tag{28}$$

Since the estimation problem $\mathcal{E}$ is assumed regular up to order 1 at $\mathbf{x}_0$, the associated RKHS $\mathcal{H}_{\mathcal{E},\mathbf{x}_0}$ is differentiable up to order 1. This differentiability is used in the proof of the following result [10, Section 4.4.2].

Theorem III.2. *Consider an estimation problem that is regular up to order 1 at a reference parameter vector $\mathbf{x}_0 \in \mathcal{X}^{\mathrm{o}}$ in the sense of Definition* II.1. *Then, the CRB in* (27) *is obtained from* (21) *by using the subspace*

$$\mathcal{U}_{\mathrm{CR}} \triangleq \operatorname{span}\big\{\{u_0(\cdot)\} \cup \{u_l(\cdot)\}_{l\in[N]}\big\}\,,$$

with the functions

$$u_0(\cdot) \triangleq R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\mathbf{x}_0) \in \mathcal{H}_{\mathcal{E},\mathbf{x}_0}\,,$$
$$u_l(\cdot) \triangleq \frac{\partial R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\mathbf{x})}{\partial x_l}\bigg|_{\mathbf{x}=\mathbf{x}_0} \in \mathcal{H}_{\mathcal{E},\mathbf{x}_0}\,,\quad l\in[N]\,.$$

C. Constrained Cramér–Rao Bound

The constrained CRB [32]–[34] is an evolution of the CRB in (27) for estimation problems $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ with a parameter set of the form

$$\mathcal{X} = \big\{\mathbf{x}\in\mathbb{R}^N \big| \mathbf{f}(\mathbf{x}) = \mathbf{0}\big\}\,, \tag{29}$$

where $\mathbf{f}(\cdot) : \mathbb{R}^N \to \mathbb{R}^Q$ with $Q \leq N$ is a continuously differentiable function. We assume that the set $\mathcal{X}$ has a nonempty interior. Moreover, we require the Jacobian matrix $\mathbf{F}(\mathbf{x}) \triangleq \frac{\partial\mathbf{f}(\mathbf{x})}{\partial\mathbf{x}} \in \mathbb{R}^{Q\times N}$ to have rank Q whenever $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, i.e., for every $\mathbf{x}\in\mathcal{X}$. This full-rank requirement implies that the constraints represented by $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ are nonredundant [33]. Under these conditions, the implicit function theorem [34, Theorem 3.3], [13, Theorem 9.28] states that for any $\mathbf{x}_0 \in \mathcal{X}$, with $\mathcal{X}$ given by (29), there exists a continuously differentiable map $\mathbf{r}(\cdot)$ from an open set $\mathcal{O} \subseteq \mathbb{R}^{N-Q}$ into a set $\mathcal{P} \subseteq \mathcal{X}$ containing $\mathbf{x}_0$, i.e.,

$$\mathbf{r}(\cdot) : \mathcal{O} \subseteq \mathbb{R}^{N-Q} \to \mathcal{P} \subseteq \mathcal{X}, \quad \text{with } \mathbf{x}_0 \in \mathcal{P}. \tag{30}$$

The constrained CRB in the form presented in [33] reads

$$v(\hat{g}(\cdot);\mathbf{x}_0) \geq \mathbf{b}^T(\mathbf{x}_0)\mathbf{U}(\mathbf{x}_0)\left(\mathbf{U}^T(\mathbf{x}_0)\,\mathbf{J}(\mathbf{x}_0)\,\mathbf{U}(\mathbf{x}_0)\right)^{\dagger} \times \mathbf{U}^T(\mathbf{x}_0)\,\mathbf{b}(\mathbf{x}_0)\,, \tag{31}$$

where $\mathbf{b}(\mathbf{x}_0) = \frac{\partial\gamma(\mathbf{x})}{\partial\mathbf{x}}\big|_{\mathbf{x}=\mathbf{x}_0}$, $\mathbf{J}(\mathbf{x}_0)$ is the Fisher information matrix defined in (28), and $\mathbf{U}(\mathbf{x}_0) \in \mathbb{R}^{N\times(N-Q)}$ is any matrix whose columns form an ONB for the null space of the Jacobian matrix $\mathbf{F}(\mathbf{x}_0)$, i.e.,

$$\mathbf{F}(\mathbf{x}_0)\,\mathbf{U}(\mathbf{x}_0) = \mathbf{0}\,,\quad \mathbf{U}^T(\mathbf{x}_0)\,\mathbf{U}(\mathbf{x}_0) = \mathbf{I}_{N-Q}\,.$$

The next result is proved in [10, Section 4.4.2].

Theorem III.3. *Consider an estimation problem that is regular up to order 1 in the sense of Definition* II.1. *Then, for a reference parameter vector $\mathbf{x}_0 \in \mathcal{X}^{\mathrm{o}}$, the constrained CRB in* (31) *is obtained from* (21) *by using the subspace*

$$\mathcal{U}_{\mathrm{CCR}} \triangleq \operatorname{span}\big\{\{u_0(\cdot)\} \cup \{u_l(\cdot)\}_{l\in[N-Q]}\big\}\,,$$

with the functions

$$u_0(\cdot) \triangleq R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\mathbf{x}_0) \in \mathcal{H}_{\mathcal{E},\mathbf{x}_0}\,,$$
$$u_l(\cdot) \triangleq \frac{\partial R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\mathbf{r}(\boldsymbol{\theta}))}{\partial\theta_l}\bigg|_{\boldsymbol{\theta}=\mathbf{r}^{-1}(\mathbf{x}_0)} \in \mathcal{H}_{\mathcal{E},\mathbf{x}_0}\,,\quad l\in[N-Q]\,,$$

where $\mathbf{r}(\cdot)$ is any continuously differentiable function of the form (30).

D. Bhattacharyya Bound

Whereas the CRB depends only on the first-order partial derivatives of $f(\mathbf{y};\mathbf{x})$ with respect to $\mathbf{x}$, the Bhattacharyya bound [35], [36] involves also higher-order derivatives. For an estimation problem $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ that is regular at $\mathbf{x}_0 \in \mathcal{X}^{\mathrm{o}}$ up to order $m\in\mathbb{N}$, the Bhattacharyya bound states that

$$v(\hat{g}(\cdot);\mathbf{x}_0) \,\geq\, \mathbf{a}^T(\mathbf{x}_0)\,\mathbf{B}^{\dagger}(\mathbf{x}_0)\,\mathbf{a}(\mathbf{x}_0)\,, \tag{32}$$

where the vector $\mathbf{a}(\mathbf{x}_0) \in \mathbb{R}^L$ and the matrix $\mathbf{B}(\mathbf{x}_0) \in \mathbb{R}^{L\times L}$ are given elementwise by $\big(\mathbf{a}(\mathbf{x}_0)\big)_l \triangleq \frac{\partial^{\mathbf{p}_l}\gamma(\mathbf{x})}{\partial\mathbf{x}^{\mathbf{p}_l}}\big|_{\mathbf{x}=\mathbf{x}_0}$ and

$$\big(\mathbf{B}(\mathbf{x}_0)\big)_{l,l'} \triangleq \mathsf{E}_{\mathbf{x}_0}\left\{\frac{1}{f^2(\mathbf{y};\mathbf{x}_0)}\,\frac{\partial^{\mathbf{p}_l} f(\mathbf{y};\mathbf{x})}{\partial\mathbf{x}^{\mathbf{p}_l}}\,\frac{\partial^{\mathbf{p}_{l'}} f(\mathbf{y};\mathbf{x})}{\partial\mathbf{x}^{\mathbf{p}_{l'}}}\bigg|_{\mathbf{x}=\mathbf{x}_0}\right\},$$

respectively. Here, the $\mathbf{p}_l$, $l\in[L]$ are L distinct multi-indices with $(\mathbf{p}_l)_k \leq m$.

The following result is proved in [10, Section 4.4.3].

Theorem III.4. *Consider an estimation problem that is regular up to order m at a reference parameter vector $\mathbf{x}_0 \in \mathcal{X}^{\mathrm{o}}$ in the sense of Definition* II.1. *Then, the Bhattacharyya bound in* (32) *is obtained from* (21) *by using the subspace*

$$\mathcal{U}_{\mathrm{B}} \triangleq \operatorname{span}\big\{\{u_0(\cdot)\} \cup \{u_l(\cdot)\}_{l\in[L]}\big\}\,,$$

with the functions

$$u_0(\cdot) \triangleq R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\mathbf{x}_0) \in \mathcal{H}_{\mathcal{E},\mathbf{x}_0}\,,$$
$$u_l(\cdot) \triangleq \frac{\partial^{\mathbf{p}_l} R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\mathbf{x})}{\partial \mathbf{x}^{\mathbf{p}_l}}\bigg|_{\mathbf{x}=\mathbf{x}_0} \in \mathcal{H}_{\mathcal{E},\mathbf{x}_0}\,, \quad l \in [L]\,. \tag{33}$$

While the RKHS interpretation of the Bhattacharyya bound has been presented previously in [3] for a specific estimation problem, the above result holds for general estimation problems. We note that the bound tends to become higher (tighter) if L is increased in the sense that additional functions $u_l(\cdot)$ are used (i.e., in addition to the functions already used). Finally, we note that the CRB subspace $\mathcal{U}_{\text{CR}}$ in Theorem III.2 is obtained as a special case of the Bhattacharyya bound subspace $\mathcal{U}_{\text{B}}$ by setting $L = N$, $m = 1$, and $\mathbf{p}_l = \mathbf{e}_l$ in (33).

E. Hammersley-Chapman-Robbins Bound

A drawback of the CRB and the Bhattacharyya bound is that they exploit only the local structure of an estimation problem $\mathcal{E}$ around a specific point $\mathbf{x}_0 \in \mathcal{X}^{\text{o}}$ [35]. As an illustrative example, consider two different estimation problems $\mathcal{E}_1 = \big(\mathcal{X}_1, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ and $\mathcal{E}_2 = \big(\mathcal{X}_2, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ with the same statistical model $f(\mathbf{y};\mathbf{x})$ and parameter function $g(\cdot)$ but different parameter sets $\mathcal{X}_1$ and $\mathcal{X}_2$. These parameter sets are assumed to be open balls centered at $\mathbf{x}_0$ with different radii r_1 and r_2, i.e., $\mathcal{X}_1 = \mathcal{B}(\mathbf{x}_0, r_1)$ and $\mathcal{X}_2 = \mathcal{B}(\mathbf{x}_0, r_2)$ with $r_1 \neq r_2$. Then the CRB at $\mathbf{x}_0$ for both estimation problems will be identical, irrespective of the values of r_1 and r_2, and similarly for the Bhattacharyya bound. Thus, these bounds do not take into account a part of the information contained in the parameter set $\mathcal{X}$. The Barankin bound, on the other hand, exploits the full information carried by the parameter set $\mathcal{X}$ since it is the tightest possible lower bound on the estimator variance. However, the Barankin bound is difficult to evaluate in general.

The Hammersley-Chapman-Robbins bound (HCRB) [37]–[39] is a lower bound on the estimator variance that takes into account the global structure of the estimation problem associated with the entire parameter set $\mathcal{X}$. It can be evaluated much more easily than the Barankin bound, and it does not require the estimation problem to be regular. Based on a suitably chosen set of "test points" $\{\mathbf{x}_1, \ldots, \mathbf{x}_L\} \subseteq \mathcal{X}$, the HCRB states that [37]

$$v(\hat{g}(\cdot);\mathbf{x}_0) \,\geq\, \mathbf{m}^T(\mathbf{x}_0)\mathbf{V}^{\dagger}(\mathbf{x}_0)\,\mathbf{m}(\mathbf{x}_0)\,, \tag{34}$$

where the vector $\mathbf{m}(\mathbf{x}_0) \in \mathbb{R}^L$ and the matrix $\mathbf{V}(\mathbf{x}_0) \in \mathbb{R}^{L\times L}$ are given elementwise by $\big(\mathbf{m}(\mathbf{x}_0)\big)_l \triangleq \gamma(\mathbf{x}_l) - \gamma(\mathbf{x}_0)$ and

$$\big(\mathbf{V}(\mathbf{x}_0)\big)_{l,l'} \triangleq \mathsf{E}_{\mathbf{x}_0}\bigg\{\frac{[f(\mathbf{y};\mathbf{x}_l) - f(\mathbf{y};\mathbf{x}_0)][f(\mathbf{y};\mathbf{x}_{l'}) - f(\mathbf{y};\mathbf{x}_0)]}{f^2(\mathbf{y};\mathbf{x}_0)}\bigg\}\,,$$

respectively.

The following result is proved in [10, Section 4.4.4].

Theorem III.5. *The HCRB in* (34)*, with test points* $\{\mathbf{x}_l\}_{l\in[L]} \subseteq \mathcal{X}$*, is obtained from* (21) *by using the subspace*

$$\mathcal{U}_{\text{HCR}} \triangleq \operatorname{span}\big\{\{u_0(\cdot)\} \cup \{u_l(\cdot)\}_{l\in[L]}\big\}\,,$$

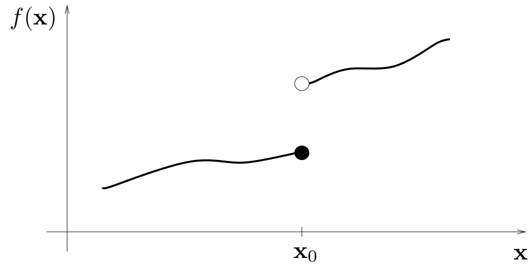

Fig. 1. Graph of a function that is lower semi-continuous at $\mathbf{x}_0$. The solid dot indicates the function value $f(\mathbf{x}_0)$.

with the functions

$$u_0(\cdot) \triangleq R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\mathbf{x}_0) \in \mathcal{H}_{\mathcal{E},\mathbf{x}_0}\,,$$
$$u_l(\cdot) \triangleq R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\mathbf{x}_l) - R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\mathbf{x}_0)\,, \quad l \in [L]\,.$$

The HCRB tends to become higher (tighter) if L is increased in the sense that test points $\mathbf{x}_l$ or, equivalently, functions $u_l(\cdot)$ are added to those already used.

F. Lower Semi-Continuity of the Barankin Bound

For a given estimation problem $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ and a prescribed bias function $c(\cdot)$, it is sometimes of interest to characterize not only the minimum achievable variance $M(c(\cdot),\mathbf{x}_0)$ at a single parameter vector $\mathbf{x}_0 \in \mathcal{X}$ but also how $M(c(\cdot),\mathbf{x}_0)$ changes if $\mathbf{x}_0$ is varied. The following result is proved in Appendix A.

Theorem III.6. *Consider an estimation problem* $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ *with parameter set* $\mathcal{X} \subseteq \mathbb{R}^N$ *and a prescribed bias function* $c(\cdot)\colon \mathcal{X} \to \mathbb{R}$ *that is valid at all* $\mathbf{x}_0 \in \mathcal{C}$ *for some open set* $\mathcal{C} \subseteq \mathcal{X}$ *and for which the associated prescribed mean function* $\gamma(\cdot) = c(\cdot) + g(\cdot)$ *is a continuous function on* $\mathcal{C}$*. Furthermore assume that for any fixed* $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, $R_{\mathcal{E},\mathbf{x}_0}(\mathbf{x}_1,\mathbf{x}_2)$ *is continuous with respect to* $\mathbf{x}_0$ *on* $\mathcal{C}$*, i.e.,*

$$\lim_{\mathbf{x}'_0 \to \mathbf{x}_0} R_{\mathcal{E},\mathbf{x}'_0}(\mathbf{x}_1,\mathbf{x}_2) = R_{\mathcal{E},\mathbf{x}_0}(\mathbf{x}_1,\mathbf{x}_2)\,, \quad \forall\, \mathbf{x}_0 \in \mathcal{C}\,,\ \forall\, \mathbf{x}_1,\mathbf{x}_2 \in \mathcal{X}. \tag{35}$$

Then, the minimum achievable variance $M(c(\cdot),\mathbf{x})$*, viewed as a function of* $\mathbf{x}$*, is lower semi-continuous on* $\mathcal{C}$*.*

A schematic illustration of a lower semi-continuous function is given in Fig. 1. The application of Theorem III.6 to the estimation problems considered in [40]—corresponding to the linear/Gaussian model with a sparse parameter vector—allows us to conclude that the "sparse CRB" introduced in [40] cannot be maximally tight, i.e., it is not equal to the minimum achievable variance. Indeed, the sparse CRB derived in [40] is in general a strictly upper semi-continuous[3] function of the parameter vector $\mathbf{x}$, whereas the minimum achievable variance $M(c(\cdot),\mathbf{x})$ is lower semi-continuous according to Theorem III.6. Since a function cannot be simultaneously lower semi-continuous and strictly upper semi-continuous, the sparse CRB cannot be equal to $M(c(\cdot),\mathbf{x})$.

[3] A function is said to be *strictly upper semi-continuous* if it is upper semi-continuous but not continuous.

IV. Sufficient Statistics

For some estimation problems $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$, the observation $\mathbf{y} \in \mathbb{R}^M$ contains information that is irrelevant to $\mathcal{E}$, and thus $\mathbf{y}$ can be compressed in some sense. Accordingly, let us replace $\mathbf{y}$ by a transformed observation $\mathbf{z} = \mathbf{t}(\mathbf{y}) \in \mathbb{R}^K$, with a deterministic mapping $\mathbf{t}(\cdot): \mathbb{R}^M \to \mathbb{R}^K$. A compression is achieved if $K < M$. Any transformed observation $\mathbf{z} = \mathbf{t}(\mathbf{y})$ is termed a *statistic*, and it is said to be a *sufficient statistic* if it preserves all the information that is relevant to $\mathcal{E}$ [1], [15]–[18], [41]. In particular, a sufficient statistic preserves the minimum achievable variance (Barankin bound) $M(c(\cdot), \mathbf{x}_0)$. In the following, the mapping $\mathbf{t}(\cdot)$ will be assumed to be measurable.

For a given reference parameter vector $\mathbf{x}_0 \in \mathcal{X}$, we consider estimation problems $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ for which there exists a dominating measure $\mu_{\mathcal{E}}$ such that the pdfs $\{f(\mathbf{y};\mathbf{x})\}_{\mathbf{x}\in\mathcal{X}}$ are well defined with respect to $\mu_{\mathcal{E}}$ and condition (5) is satisfied. The Neyman-Fisher factorization theorem [15]–[18] then states that the statistic $\mathbf{z} = \mathbf{t}(\mathbf{y})$ is sufficient for $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ if and only if $f(\mathbf{y};\mathbf{x})$ can be factored as

$$f(\mathbf{y};\mathbf{x}) = h(\mathbf{t}(\mathbf{y});\mathbf{x})\, k(\mathbf{y})\,, \tag{36}$$

where $h(\cdot\,;\mathbf{x})$ and $k(\cdot)$ are nonnegative functions and the function $k(\cdot)$ does not depend on $\mathbf{x}$. Relation (36) has to be satisfied for every $\mathbf{y} \in \mathbb{R}^M$ except for a set of measure zero with respect to the dominating measure $\mu_{\mathcal{E}}$.

The probability measure on $\mathbb{R}^K$ (equipped with the system of K-dimensional Borel sets, cf. [1, Section 10]) that is induced by the random vector $\mathbf{z} = \mathbf{t}(\mathbf{y})$ is obtained as $\mu^{\mathbf{z}}_{\mathbf{x}} = \mu^{\mathbf{y}}_{\mathbf{x}}\mathbf{t}^{-1}$ [16], [17]. According to Section II-B1, under condition (5), the measure $\mu^{\mathbf{y}}_{\mathbf{x}_0}$ dominates the measures $\{\mu^{\mathbf{y}}_{\mathbf{x}}\}_{\mathbf{x}\in\mathcal{X}}$. This, in turn, implies via [16, Lemma 4] that the measure $\mu^{\mathbf{z}}_{\mathbf{x}_0}$ dominates the measures $\{\mu^{\mathbf{z}}_{\mathbf{x}}\}_{\mathbf{x}\in\mathcal{X}}$, and therefore that, for each $\mathbf{x} \in \mathcal{X}$, there exists a pdf $f(\mathbf{z};\mathbf{x})$ with respect to the measure $\mu^{\mathbf{z}}_{\mathbf{x}_0}$. This pdf is given by the following result. (Note that we do not assume condition (8).)

Lemma IV.1. *Consider an estimation problem $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ satisfying* (5)*, i.e., which is such that the Radon-Nikodym derivative of $\mu^{\mathbf{y}}_{\mathbf{x}}$ with respect to $\mu^{\mathbf{y}}_{\mathbf{x}_0}$ is well defined and given by the likelihood ratio $\rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x})$. Furthermore consider a sufficient statistic $\mathbf{z} = \mathbf{t}(\mathbf{y})$ for $\mathcal{E}$. Then, the pdf of $\mathbf{z}$ with respect to the dominating measure $\mu^{\mathbf{z}}_{\mathbf{x}_0}$ is given by*

$$f(\mathbf{z};\mathbf{x}) = \frac{h(\mathbf{z};\mathbf{x})}{h(\mathbf{z};\mathbf{x}_0)}\,, \tag{37}$$

where the function $h(\mathbf{z};\mathbf{x})$ is obtained from the factorization (36).

Proof: The pdf $f(\mathbf{z};\mathbf{x})$ of $\mathbf{z}$ with respect to $\mu^{\mathbf{z}}_{\mathbf{x}_0}$ is defined by the relation

$$\mathsf{E}_{\mathbf{x}_0}\big\{I_{\mathcal{A}}(\mathbf{z}) f(\mathbf{z};\mathbf{x})\big\} = \mathrm{P}_{\mathbf{x}}\{\mathbf{z}\in\mathcal{A}\}\,, \tag{38}$$

which has to be satisfied for every measurable set $\mathcal{A} \subseteq \mathbb{R}^K$ [1]. Denoting the pre-image of $\mathcal{A}$ under the mapping $\mathbf{t}(\cdot)$ by $\mathbf{t}^{-1}(\mathcal{A}) \triangleq \big\{\mathbf{y}\big|\mathbf{t}(\mathbf{y})\in\mathcal{A}\big\} \subseteq \mathbb{R}^M$, we have

$$\begin{aligned}
\mathsf{E}_{\mathbf{x}_0}\bigg\{I_{\mathcal{A}}(\mathbf{z})\,\frac{h(\mathbf{z};\mathbf{x})}{h(\mathbf{z};\mathbf{x}_0)}\bigg\} &\stackrel{(a)}{=} \mathsf{E}_{\mathbf{x}_0}\bigg\{I_{\mathcal{A}}(\mathbf{t}(\mathbf{y}))\,\frac{h(\mathbf{t}(\mathbf{y});\mathbf{x})}{h(\mathbf{t}(\mathbf{y});\mathbf{x}_0)}\bigg\}\\
&= \mathsf{E}_{\mathbf{x}_0}\bigg\{I_{\mathbf{t}^{-1}(\mathcal{A})}(\mathbf{y})\,\frac{h(\mathbf{t}(\mathbf{y});\mathbf{x})}{h(\mathbf{t}(\mathbf{y});\mathbf{x}_0)}\bigg\}\\
&\stackrel{(36),(6)}{=} \mathsf{E}_{\mathbf{x}_0}\Big\{I_{\mathbf{t}^{-1}(\mathcal{A})}(\mathbf{y})\,\rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x})\Big\}\\
&\stackrel{(b)}{=} \mathrm{P}_{\mathbf{x}}\{\mathbf{y}\in\mathbf{t}^{-1}(\mathcal{A})\}\\
&= \mathrm{P}_{\mathbf{x}}\{\mathbf{z}\in\mathcal{A}\}\,,
\end{aligned} \tag{39}$$

where step (a) follows from [1, Theorem 16.12] and (b) is due to the fact that the Radon-Nikodym derivative of $\mu^{\mathbf{y}}_{\mathbf{x}}$ with respect to $\mu^{\mathbf{y}}_{\mathbf{x}_0}$ is given by $\rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x})$ (cf. (6)), as explained in Section II-B1. Comparing (39) with (38), we conclude that $\frac{h(\mathbf{z};\mathbf{x})}{h(\mathbf{z};\mathbf{x}_0)} = f(\mathbf{z};\mathbf{x})$ up to differences on a set of measure zero (with respect to $\mu^{\mathbf{z}}_{\mathbf{x}_0}$). Note that because we require $\mathbf{t}(\cdot)$ to be a measurable mapping, it is guaranteed that the set $\mathbf{t}^{-1}(\mathcal{A}) = \big\{\mathbf{y}\big|\mathbf{t}(\mathbf{y})\in\mathcal{A}\big\}$ is measurable for any measurable set $\mathcal{A} \subseteq \mathbb{R}^K$. □

Consider next an estimation problem $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ satisfying (8), so that the kernel $R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\cdot)$ exists according to (9). Let $\mathbf{z} = \mathbf{t}(\mathbf{y})$ be a sufficient statistic. We can then define the modified estimation problem $\mathcal{E}' \triangleq \big(\mathcal{X}, f(\mathbf{z};\mathbf{x}), g(\cdot)\big)$, which is based on the observation $\mathbf{z}$ and whose statistical model is given by the pdf $f(\mathbf{z};\mathbf{x})$ (cf. (37)). The following theorem states that the RKHS associated with $\mathcal{E}'$ equals the RKHS associated with $\mathcal{E}$.

Theorem IV.2. *Consider an estimation problem $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ satisfying* (8) *and a reference parameter vector $\mathbf{x}_0 \in \mathcal{X}$. For a sufficient statistic $\mathbf{z} = \mathbf{t}(\mathbf{y})$, consider the modified estimation problem $\mathcal{E}' = \big(\mathcal{X}, f(\mathbf{z};\mathbf{x}), g(\cdot)\big)$. Then, $\mathcal{E}'$ also satisfies* (8) *and furthermore $R_{\mathcal{E}',\mathbf{x}_0}(\cdot\,,\cdot) = R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\cdot)$ and $\mathcal{H}_{\mathcal{E}',\mathbf{x}_0} = \mathcal{H}_{\mathcal{E},\mathbf{x}_0}$.*

Proof: We have

$$\begin{aligned}
R_{\mathcal{E},\mathbf{x}_0}(\mathbf{x}_1,\mathbf{x}_2) &\stackrel{(9),(7)}{=} \mathsf{E}_{\mathbf{x}_0}\big\{\rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x}_1)\rho_{\mathcal{E},\mathbf{x}_0}(\mathbf{y},\mathbf{x}_2)\big\}\\
&\stackrel{(36),(6)}{=} \mathsf{E}_{\mathbf{x}_0}\bigg\{\frac{h(\mathbf{t}(\mathbf{y});\mathbf{x}_1)h(\mathbf{t}(\mathbf{y});\mathbf{x}_2)}{h^2(\mathbf{t}(\mathbf{y});\mathbf{x}_0)}\bigg\}\\
&\stackrel{(a)}{=} \mathsf{E}_{\mathbf{x}_0}\bigg\{\frac{h(\mathbf{z};\mathbf{x}_1)h(\mathbf{z};\mathbf{x}_2)}{h^2(\mathbf{z};\mathbf{x}_0)}\bigg\}\\
&\stackrel{(37)}{=} \mathsf{E}_{\mathbf{x}_0}\big\{f(\mathbf{z};\mathbf{x}_1)f(\mathbf{z};\mathbf{x}_2)\big\}\\
&\stackrel{(37)}{=} \mathsf{E}_{\mathbf{x}_0}\bigg\{\frac{f(\mathbf{z};\mathbf{x}_1)f(\mathbf{z};\mathbf{x}_2)}{f^2(\mathbf{z};\mathbf{x}_0)}\bigg\}\\
&\stackrel{(7),(9)}{=} R_{\mathcal{E}',\mathbf{x}_0}(\mathbf{x}_1,\mathbf{x}_2)\,,
\end{aligned} \tag{40}$$

where, as before, step (a) follows from [1, Theorem 16.12]. From (40), we conclude that if $\mathcal{E}$ satisfies (8) then so does $\mathcal{E}'$. Moreover, from $R_{\mathcal{E}',\mathbf{x}_0}(\cdot\,,\cdot) = R_{\mathcal{E},\mathbf{x}_0}(\cdot\,,\cdot)$ in (40), it follows that $\mathcal{H}_{\mathcal{E}',\mathbf{x}_0} = \mathcal{H}(R_{\mathcal{E}',\mathbf{x}_0})$ equals $\mathcal{H}_{\mathcal{E},\mathbf{x}_0} = \mathcal{H}(R_{\mathcal{E},\mathbf{x}_0})$. □

Intuitively, one might expect that the RKHS associated with a sufficient statistic should be typically "smaller" or "simpler" than the RKHS associated with the original observation, since in general the sufficient statistic is a compressed and "more concise" version of the observation. However, Theorem IV.2 states that the RKHS remains unchanged by this compression. One possible interpretation of this fact is that the RKHS description of an estimation problem is already "maximally efficient" in the sense that it cannot be reduced or simplified by using a compressed (yet sufficiently informative) observation.

V. MVE for the Exponential Family

An important class of estimation problems is defined by statistical models belonging to an exponential family. Such models are of considerable interest in the context of MVE because, under mild conditions, the existence of a UMV estimator is guaranteed. Furthermore, any estimation problem that admits an *efficient estimator*, i.e., an estimator whose variance attains the CRB, must be necessarily based on an exponential family [15, Theorem 5.12]. In this section, we will characterize the RKHS for this class and use it to derive lower variance bounds.

A. Review of the Exponential Family

An exponential family is defined as the following parametrized set of pdfs $\{f(\mathbf{y};\mathbf{x})\}_{\mathbf{x}\in\mathcal{X}}$ (with respect to the Lebesgue measure on $\mathbb{R}^M$) [15], [42], [43]:

$$f(\mathbf{y};\mathbf{x}) \,=\, \exp\!\big(\boldsymbol{\phi}^T(\mathbf{y})\mathbf{u}(\mathbf{x}) - A(\mathbf{x})\big)\, h(\mathbf{y})\,,$$

with the *sufficient statistic* $\boldsymbol{\phi}(\cdot):\mathbb{R}^M\to\mathbb{R}^Q$, the *parameter function* $\mathbf{u}(\cdot):\mathbb{R}^N\to\mathbb{R}^Q$, the *cumulant function* $A(\cdot):\mathbb{R}^N\to\mathbb{R}$, and the *weight function* $h(\cdot):\mathbb{R}^M\to\mathbb{R}$. Many well-known statistical models are special instances of an exponential family [43]. Without loss of generality, we can restrict ourselves to an exponential family in *canonical form* [15], for which $Q=N$ and $\mathbf{u}(\mathbf{x})=\mathbf{x}$, i.e.,

$$f^{(A)}(\mathbf{y};\mathbf{x}) \,=\, \exp\!\big(\boldsymbol{\phi}^T(\mathbf{y})\mathbf{x} - A(\mathbf{x})\big)\, h(\mathbf{y})\,. \tag{41}$$

Here, the superscript $^{(A)}$ emphasizes the importance of the cumulant function $A(\cdot)$ in the characterization of an exponential family. In what follows, we assume that the parameter space is chosen as $\mathcal{X}\subseteq\mathcal{N}$, where $\mathcal{N}\subseteq\mathbb{R}^N$ is the *natural parameter space* defined as

$$\mathcal{N} \,\triangleq\, \bigg\{\mathbf{x}\in\mathbb{R}^N \,\bigg|\, \int_{\mathbb{R}^M}\exp\!\big(\boldsymbol{\phi}^T(\mathbf{y})\mathbf{x}\big)\,h(\mathbf{y})\,d\mathbf{y}<\infty\bigg\}\,.$$

From the normalization constraint $\int_{\mathbb{R}^M} f^{(A)}(\mathbf{y};\mathbf{x})\,d\mathbf{y}=1$, it follows that the cumulant function $A(\cdot)$ is determined by the sufficient statistic $\boldsymbol{\phi}(\cdot)$ and the weight function $h(\cdot)$ as

$$A(\mathbf{x}) \,=\, \log\!\bigg(\int_{\mathbb{R}^M}\exp\!\big(\boldsymbol{\phi}^T(\mathbf{y})\mathbf{x}\big)\,h(\mathbf{y})\,d\mathbf{y}\bigg), \qquad \mathbf{x}\in\mathcal{N}\,.$$

The *moment-generating function* of $f^{(A)}(\mathbf{y};\mathbf{x})$ is defined as

$$\lambda(\mathbf{x}) \,\triangleq\, \exp(A(\mathbf{x})) \,=\, \int_{\mathbb{R}^M}\exp\!\big(\boldsymbol{\phi}^T(\mathbf{y})\mathbf{x}\big)\,h(\mathbf{y})\,d\mathbf{y}\,, \quad \mathbf{x}\in\mathcal{N}. \tag{42}$$

Note that

$$\mathcal{N} = \big\{\mathbf{x}\in\mathbb{R}^N \,\big|\, \lambda(\mathbf{x})<\infty\big\}\,.$$

Assuming a random vector $\mathbf{y}\sim f^{(A)}(\mathbf{y};\mathbf{x})$, it is known [42, Theorem 2.2], [43, Proposition 3.1] that for any $\mathbf{x}\in\mathcal{X}^{\mathrm{o}}$ and $\mathbf{p}\in\mathbb{Z}_+^N$, the moments $\mathsf{E}_{\mathbf{x}}\{\boldsymbol{\phi}^{\mathbf{p}}(\mathbf{y})\}$ exist, i.e., $\mathsf{E}_{\mathbf{x}}\{\boldsymbol{\phi}^{\mathbf{p}}(\mathbf{y})\}<\infty$, and they can be calculated from the partial derivatives of $\lambda(\mathbf{x})$ according to

$$\mathsf{E}_{\mathbf{x}}\{\boldsymbol{\phi}^{\mathbf{p}}(\mathbf{y})\} \,=\, \frac{1}{\lambda(\mathbf{x})}\frac{\partial^{\mathbf{p}}\lambda(\mathbf{x})}{\partial\mathbf{x}^{\mathbf{p}}}\,. \tag{43}$$

Thus, the partial derivatives $\frac{\partial^{\mathbf{p}}\lambda(\mathbf{x})}{\partial\mathbf{x}^{\mathbf{p}}}$ exist for any $\mathbf{x}\in\mathcal{X}^{\mathrm{o}}$ and $\mathbf{p}\in\mathbb{Z}_+^N$. Moreover, they depend continuously on $\mathbf{x}\in\mathcal{X}^{\mathrm{o}}$ [42], [43].

B. RKHS Associated with an Exponential Family Based MVP

Consider an estimation problem $\mathcal{E}^{(A)}\triangleq\big(\mathcal{X},f^{(A)}(\mathbf{y};\mathbf{x}),g(\cdot)\big)$ with an exponential family statistical model $\{f^{(A)}(\mathbf{y};\mathbf{x})\}_{\mathbf{x}\in\mathcal{X}}$ as defined in (41), and a fixed $\mathbf{x}_0\in\mathcal{X}$. Consider further the RKHS $\mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}$. Its kernel is obtained as

$$\begin{aligned}
&R_{\mathcal{E}^{(A)},\mathbf{x}_0}(\mathbf{x}_1,\mathbf{x}_2)\\
&\overset{(9)}{=}\ \mathsf{E}_{\mathbf{x}_0}\bigg\{\frac{f^{(A)}(\mathbf{y};\mathbf{x}_1)f^{(A)}(\mathbf{y};\mathbf{x}_2)}{(f^{(A)}(\mathbf{y};\mathbf{x}_0))^2}\bigg\} &(44)\\
&\overset{(41)}{=}\ \mathsf{E}_{\mathbf{x}_0}\bigg\{\frac{\exp\!\big(\boldsymbol{\phi}^T(\mathbf{y})\mathbf{x}_1-A(\mathbf{x}_1)\big)\exp\!\big(\boldsymbol{\phi}^T(\mathbf{y})\mathbf{x}_2-A(\mathbf{x}_2)\big)}{\exp\!\big(2\big[\boldsymbol{\phi}^T(\mathbf{y})\mathbf{x}_0-A(\mathbf{x}_0)\big]\big)}\bigg\}\\
&=\ \mathsf{E}_{\mathbf{x}_0}\big\{\exp\!\big(\boldsymbol{\phi}^T(\mathbf{y})\,(\mathbf{x}_1+\mathbf{x}_2-2\mathbf{x}_0)-A(\mathbf{x}_1)-A(\mathbf{x}_2)\\
&\qquad\qquad +2A(\mathbf{x}_0)\big)\big\}\\
&\overset{(41)}{=}\ \exp\!\big(-A(\mathbf{x}_1)-A(\mathbf{x}_2)+2A(\mathbf{x}_0)\big)\\
&\qquad\times\int_{\mathbb{R}^M}\exp\!\big(\boldsymbol{\phi}^T(\mathbf{y})\,(\mathbf{x}_1+\mathbf{x}_2-2\mathbf{x}_0)\big)\\
&\qquad\qquad\times\exp\!\big(\boldsymbol{\phi}^T(\mathbf{y})\mathbf{x}_0-A(\mathbf{x}_0)\big)\,h(\mathbf{y})\,d\mathbf{y}\\
&=\ \exp\!\big(-A(\mathbf{x}_1)-A(\mathbf{x}_2)+A(\mathbf{x}_0)\big)\\
&\qquad\times\int_{\mathbb{R}^M}\exp\!\big(\boldsymbol{\phi}^T(\mathbf{y})\,(\mathbf{x}_1+\mathbf{x}_2-\mathbf{x}_0)\big)\,h(\mathbf{y})\,d\mathbf{y}\\
&\overset{(42)}{=}\ \frac{\lambda(\mathbf{x}_1+\mathbf{x}_2-\mathbf{x}_0)\,\lambda(\mathbf{x}_0)}{\lambda(\mathbf{x}_1)\,\lambda(\mathbf{x}_2)}\,. &(45)
\end{aligned}$$

Because (44) and (45) are equal, we see that condition (8) is satisfied, i.e., $\mathsf{E}_{\mathbf{x}_0}\Big\{\frac{f^{(A)}(\mathbf{y};\mathbf{x}_1)f^{(A)}(\mathbf{y};\mathbf{x}_2)}{(f^{(A)}(\mathbf{y};\mathbf{x}_0))^2}\Big\}<\infty$ for all $\mathbf{x}_1,\mathbf{x}_2\in\mathcal{X}$, if and only if $\frac{\lambda(\mathbf{x}_1+\mathbf{x}_2-\mathbf{x}_0)\,\lambda(\mathbf{x}_0)}{\lambda(\mathbf{x}_1)\,\lambda(\mathbf{x}_2)}<\infty$ for all $\mathbf{x}_1,\mathbf{x}_2\in\mathcal{X}$. Since $\mathbf{x}_0\in\mathcal{X}\subseteq\mathcal{N}$, we have $\lambda(\mathbf{x}_0)<\infty$. Furthermore, $\lambda(\mathbf{x})\neq0$ for all $\mathbf{x}\in\mathcal{X}$. Therefore, (8) is satisfied if and only if $\lambda(\mathbf{x}_1+\mathbf{x}_2-\mathbf{x}_0)<\infty$. We conclude that for an estimation problem whose statistical model belongs to an exponential family, condition (8) is equivalent to

$$\mathbf{x}_1,\mathbf{x}_2\in\mathcal{X} \ \ \Rightarrow\ \ \mathbf{x}_1+\mathbf{x}_2-\mathbf{x}_0\in\mathcal{N}\,. \tag{46}$$

Furthermore, from (45) and the fact that the partial derivatives $\frac{\partial^{\mathbf{p}}\lambda(\mathbf{x})}{\partial\mathbf{x}^{\mathbf{p}}}$ exist for any $\mathbf{x}\in\mathcal{X}^{\mathrm{o}}$ and $\mathbf{p}\in\mathbb{Z}_+^N$ and depend continuously on $\mathbf{x}\in\mathcal{X}^{\mathrm{o}}$, we can conclude that the RKHS $\mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}$ is differentiable up to any order. We summarize this finding in the following lemma.

Lemma V.1. *Consider an estimation problem $\mathcal{E}^{(A)}=\big(\mathcal{X},f^{(A)}(\mathbf{y};\mathbf{x}),g(\cdot)\big)$ associated with an exponential family (cf.* (41)*) with natural parameter space $\mathcal{N}$. We assume that $\mathcal{X}\subseteq\mathcal{N}$ and that $\mathcal{X}$ satisfies condition* (46) *for some reference parameter vector $\mathbf{x}_0\in\mathcal{X}$. Then, the kernel $R_{\mathcal{E}^{(A)},\mathbf{x}_0}(\mathbf{x}_1,\mathbf{x}_2)$ and the RKHS $\mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}$ are differentiable up to any order m.*

Next, by combining Lemma V.1 with (20), we will derive simple lower bounds on the variance of estimators with a prescribed bias function.

C. Variance Bounds for the Exponential Family

If $\mathcal{X}^{o}$ is nonempty, the sufficient statistic $\boldsymbol{\phi}(\cdot)$ is a *complete* sufficient statistic for the estimation problem $\mathcal{E}^{(A)}$, and thus there exists a UMV estimator $\hat{g}_{\text{UMV}}(\cdot)$ for any valid bias function $c(\cdot)$ [15, p. 42]. This UMV estimator is given by the conditional expectation[4]

$$\hat{g}_{\text{UMV}}(\mathbf{y}) = \mathsf{E}_{\mathbf{x}}\{\hat{g}_0(\mathbf{y})|\boldsymbol{\phi}(\mathbf{y})\}\,, \tag{47}$$

where $\hat{g}_0(\cdot)$ is any estimator with bias function $c(\cdot)$, i.e., $b(\hat{g}_0(\cdot);\mathbf{x}_0) = c(\mathbf{x})$ for all $\mathbf{x}\in\mathcal{X}$. The minimum achievable variance $M(c(\cdot),\mathbf{x}_0)$ is then equal to the variance of $\hat{g}_{\text{UMV}}(\cdot)$ at $\mathbf{x}_0$, i.e., $M(c(\cdot),\mathbf{x}_0) = v(\hat{g}_{\text{UMV}}(\cdot);\mathbf{x}_0)$ [15, p. 89]. However, it may be difficult to actually construct the UMV estimator via (47) and to calculate its variance. In fact, it may be already a difficult task to find an estimator $\hat{g}_0(\cdot)$ whose bias function equals $c(\cdot)$. Therefore, it is still of interest to find simple closed-form lower bounds on the variance of any estimator with bias $c(\cdot)$.

Theorem V.2. *Consider an estimation problem $\mathcal{E}^{(A)} = \big(\mathcal{X}, f^{(A)}(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ with parameter set $\mathcal{X}\subseteq\mathcal{N}$ satisfying (46) and a finite set of multi-indices $\{\mathbf{p}_l\}_{l\in[L]}\subseteq\mathbb{Z}_+^N$. Then, at any $\mathbf{x}_0\in\mathcal{X}^{o}$, the variance of any estimator $\hat{g}(\cdot)$ with mean function $\gamma(\mathbf{x}) = \mathsf{E}_{\mathbf{x}}\{\hat{g}(\mathbf{y})\}$ and finite variance at $\mathbf{x}_0$ is lower bounded as*

$$v(\hat{g}(\cdot);\mathbf{x}_0) \,\geq\, \mathbf{n}^T(\mathbf{x}_0)\,\mathbf{S}^{\dagger}(\mathbf{x}_0)\,\mathbf{n}(\mathbf{x}_0) - \gamma^2(\mathbf{x}_0)\,, \tag{48}$$

where the vector $\mathbf{n}(\mathbf{x}_0)\in\mathbb{R}^L$ and the matrix $\mathbf{S}(\mathbf{x}_0)\in\mathbb{R}^{L\times L}$ are given elementwise by

$$\big(\mathbf{n}(\mathbf{x}_0)\big)_l \triangleq \sum_{\mathbf{p}\leq\mathbf{p}_l}\binom{\mathbf{p}_l}{\mathbf{p}}\,\mathsf{E}_{\mathbf{x}_0}\{\boldsymbol{\phi}^{\mathbf{p}_l-\mathbf{p}}(\mathbf{y})\}\,\frac{\partial^{\mathbf{p}}\gamma(\mathbf{x})}{\partial\mathbf{x}^{\mathbf{p}}}\bigg|_{\mathbf{x}=\mathbf{x}_0} \tag{49}$$

$$\big(\mathbf{S}(\mathbf{x}_0)\big)_{l,l'} \triangleq \mathsf{E}_{\mathbf{x}_0}\big\{\boldsymbol{\phi}^{\mathbf{p}_l+\mathbf{p}_{l'}}(\mathbf{y})\big\}\,, \tag{50}$$

respectively. Here, $\sum_{\mathbf{p}\leq\mathbf{p}_l}$ denotes the sum over all multi-indices $\mathbf{p}\in\mathbb{Z}_+^N$ such that $p_k\leq(\mathbf{p}_l)_k$ for $k\in[N]$, and $\binom{\mathbf{p}_l}{\mathbf{p}}\triangleq\prod_{k=1}^{N}\binom{(\mathbf{p}_l)_k}{p_k}$.

A proof of this result is provided in Appendix B. This proof shows that the bound (48) is obtained by projecting an appropriately transformed version of the mean function $\gamma(\cdot)$ onto the finite-dimensional subspace $\mathcal{U} = \operatorname{span}\big\{r_{\mathbf{x}_0}^{(\mathbf{p}_l)}(\cdot)\big\}_{l\in[L]}$ of an appropriately defined RKHS $\mathcal{H}(R)$, with the functions $r_{\mathbf{x}_0}^{(\mathbf{p}_l)}(\cdot)$ given by (19). If we increase the set $\big\{r_{\mathbf{x}_0}^{(\mathbf{p}_l)}(\cdot)\big\}_{l\in[L]}$ by adding further functions $r_{\mathbf{x}_0}^{(\mathbf{p}')}(\cdot)$ with multi-indices $\mathbf{p}'\notin\{\mathbf{p}_l\}_{l\in[L]}$, the subspace tends to become higher-dimensional and in turn the lower bound (48) becomes higher, i.e., tighter.

The requirement of a finite variance $v(\hat{g}(\cdot);\mathbf{x}_0)$ in Theorem V.2 implies via (10) that $\gamma(\cdot)\in\mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}$. This, in turn, guarantees via (20)—which can be invoked since due to Lemma V.1 the RKHS $\mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}$ is differentiable up to any order at $\mathbf{x}_0$—the existence of the partial derivatives $\frac{\partial^{\mathbf{p}}\gamma(\mathbf{x})}{\partial\mathbf{x}^{\mathbf{p}}}\big|_{\mathbf{x}=\mathbf{x}_0}$. Note also that the bound (48) depends on the mean function $\gamma(\cdot)$ only via the local behavior of $\gamma(\cdot)$ as given by the partial derivatives of $\gamma(\cdot)$ at $\mathbf{x}_0$ up to a suitable order.

Evaluating the bound (48) requires computation of the moments $\mathsf{E}_{\mathbf{x}_0}\big\{\boldsymbol{\phi}^{\mathbf{p}}(\mathbf{y})\big\}$. This can be done by means of message passing algorithms [43].

For the choice $L=N$ and $\mathbf{p}_l=\mathbf{e}_l$, the bound (48) is closely related to the CRB obtained for the estimation problem $\mathcal{E}^{(A)}$. In fact, the CRB for $\mathcal{E}^{(A)}$ is obtained as [15, Theorem 2.6.2]

$$v(\hat{g}(\cdot);\mathbf{x}_0) \,\geq\, \mathbf{n}^T(\mathbf{x}_0)\,\mathbf{J}^{\dagger}(\mathbf{x}_0)\,\mathbf{n}(\mathbf{x}_0)\,, \tag{51}$$

with $\big(\mathbf{n}(\mathbf{x}_0)\big)_l = \frac{\partial\gamma(\mathbf{x})}{\partial x_l}\big|_{\mathbf{x}=\mathbf{x}_0}$ and the Fisher information matrix given by

$$\mathbf{J}(\mathbf{x}_0) \,=\, \mathsf{E}_{\mathbf{x}_0}\big\{\big(\boldsymbol{\phi}(\mathbf{y})-\mathsf{E}_{\mathbf{x}_0}\{\boldsymbol{\phi}(\mathbf{y})\}\big)\big(\boldsymbol{\phi}(\mathbf{y})-\mathsf{E}_{\mathbf{x}_0}\{\boldsymbol{\phi}(\mathbf{y})\}\big)^T\big\}\,,$$

i.e., the covariance matrix of the sufficient statistic vector $\boldsymbol{\phi}(\mathbf{y})$. On the other hand, evaluating the bound (48) for $L=N$ and $\mathbf{p}_l=\mathbf{e}_l$ and assuming without loss of generality[5] that $\gamma(\mathbf{x}_0)=0$, we obtain

$$v(\hat{g}(\cdot);\mathbf{x}_0) \,\geq\, \mathbf{n}^T(\mathbf{x}_0)\,\mathbf{S}^{\dagger}(\mathbf{x}_0)\,\mathbf{n}(\mathbf{x}_0)\,, \tag{52}$$

with $\mathbf{n}(\mathbf{x}_0)$ as before and

$$\mathbf{S}(\mathbf{x}_0) \,=\, \mathsf{E}_{\mathbf{x}_0}\big\{\boldsymbol{\phi}(\mathbf{y})\boldsymbol{\phi}^T(\mathbf{y})\big\}\,.$$

Thus, the only difference is that the CRB in (51) involves the covariance matrix of the sufficient statistic $\boldsymbol{\phi}(\mathbf{y})$ whereas the bound in (52) involves the correlation matrix of $\boldsymbol{\phi}(\mathbf{y})$.

D. Reducing the Parameter Set

Using the RKHS framework, we will now show that, under mild conditions, the minimum achievable variance $M(c(\cdot),\mathbf{x}_0)$ for an exponential family type estimation problem $\mathcal{E}^{(A)} = \big(\mathcal{X}, f^{(A)}(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ is invariant to reductions of the parameter set $\mathcal{X}$. Consider two estimation problems $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ and $\mathcal{E}' = \big(\mathcal{X}', f(\mathbf{y};\mathbf{x}), g(\cdot)\big|_{\mathcal{X}'}\big)$—for now, not necessarily of the exponential family type—that differ only in their parameter sets $\mathcal{X}$ and $\mathcal{X}'$. More specifically, $\mathcal{E}'$ is obtained from $\mathcal{E}$ by reducing the parameter set, i.e., $\mathcal{X}'\subseteq\mathcal{X}$. For these two estimation problems, we consider corresponding MVPs at a specific parameter vector $\mathbf{x}_0\in\mathcal{X}'$ and for a certain prescribed bias function $c(\cdot)$. More precisely, $c(\cdot)$ is the prescribed bias function for $\mathcal{E}$ on the set $\mathcal{X}$, while the prescribed bias function for $\mathcal{E}'$ is the restriction of $c(\cdot)$ to $\mathcal{X}'$, $c(\cdot)\big|_{\mathcal{X}'}$. We will denote the minimum achievable variances of the MVPs corresponding to $\mathcal{E}$ and $\mathcal{E}'$ by $M(c(\cdot),\mathbf{x}_0)$ and $M'\big(c(\cdot)\big|_{\mathcal{X}'},\mathbf{x}_0\big)$, respectively. From (23), it follows that $M'\big(c(\cdot)\big|_{\mathcal{X}'},\mathbf{x}_0\big)\leq M(c(\cdot),\mathbf{x}_0)$, since taking the supremum over a reduced set can never result in an increase of the supremum.

The effect that a reduction of the parameter set $\mathcal{X}$ has on the minimum achievable variance can be analyzed conveniently

[4]The conditional expectation in (47) can be taken with respect to the measure $\mu_{\mathbf{x}}^{\mathbf{y}}$ for an arbitrary $\mathbf{x}\in\mathcal{X}$. Indeed, since $\boldsymbol{\phi}(\cdot)$ is a sufficient statistic, $\mathsf{E}_{\mathbf{x}}\{\hat{g}_0(\mathbf{y})|\boldsymbol{\phi}(\mathbf{y})\}$ yields the same result for every $\mathbf{x}\in\mathcal{X}$.

[5]Indeed, as we observed after Lemma III.1, the minimum achievable variance at $\mathbf{x}_0$ for prescribed mean function $\gamma(\cdot)$ is equal to the minimum achievable variance at $\mathbf{x}_0$ for prescribed mean function $\tilde{\gamma}(\mathbf{x}) = \gamma(\mathbf{x})-\gamma(\mathbf{x}_0)$. Note that $\tilde{\gamma}(\mathbf{x}_0)=0$.

within the RKHS framework. This is based on the following result [21]: Consider an RKHS $\mathcal{H}(R_1)$ of functions $f(\cdot) : \mathcal{D}_1 \to \mathbb{R}$, with kernel $R_1(\cdot,\cdot) : \mathcal{D}_1 \times \mathcal{D}_1 \to \mathbb{R}$. Let $\mathcal{D}_2 \subseteq \mathcal{D}_1$. Then, the set of functions $\big\{\tilde{f}(\cdot) \triangleq f(\cdot)\big|_{\mathcal{D}_2} \big| f(\cdot) \in \mathcal{H}(R_1)\big\}$ that is obtained by restricting each function $f(\cdot) \in \mathcal{H}(R_1)$ to the subdomain $\mathcal{D}_2$ coincides with the RKHS $\mathcal{H}(R_2)$ whose kernel $R_2(\cdot,\cdot) : \mathcal{D}_2 \times \mathcal{D}_2 \to \mathbb{R}$ is the restriction of $R_1(\cdot,\cdot) : \mathcal{D}_1 \times \mathcal{D}_1 \to \mathbb{R}$ to the subdomain $\mathcal{D}_2 \times \mathcal{D}_2$, i.e.,

$$\mathcal{H}(R_2) = \big\{\tilde{f}(\cdot) \triangleq f(\cdot)\big|_{\mathcal{D}_2} \big| f(\cdot) \in \mathcal{H}(R_1)\big\}, \\ \text{with } R_2(\cdot,\cdot) \triangleq R_1(\cdot,\cdot)\big|_{\mathcal{D}_2 \times \mathcal{D}_2}. \tag{53}$$

Furthermore, the norm of an element $\tilde{f}(\cdot) \in \mathcal{H}(R_2)$ is equal to the minimum of the norms of all functions $f(\cdot) \in \mathcal{H}(R_1)$ that coincide with $\tilde{f}(\cdot)$ on $\mathcal{D}_2$, i.e.,

$$\|\tilde{f}(\cdot)\|_{\mathcal{H}(R_2)} = \min_{\substack{f(\cdot) \in \mathcal{H}(R_1) \\ f(\cdot)|_{\mathcal{D}_2} = \tilde{f}(\cdot)}} \|f(\cdot)\|_{\mathcal{H}(R_1)}. \tag{54}$$

Consider an arbitrary but fixed $f(\cdot) \in \mathcal{H}(R_1)$, and let $\tilde{f}(\cdot) \triangleq f(\cdot)\big|_{\mathcal{D}_2}$. Because $\tilde{f}(\cdot) \in \mathcal{H}(R_2)$, we can calculate $\|\tilde{f}(\cdot)\|_{\mathcal{H}(R_2)}$. From (54), we obtain for $\|\tilde{f}(\cdot)\|_{\mathcal{H}(R_2)} = \big\|f(\cdot)\big|_{\mathcal{D}_2}\big\|_{\mathcal{H}(R_2)}$ the inequality

$$\big\|f(\cdot)\big|_{\mathcal{D}_2}\big\|_{\mathcal{H}(R_2)} \leq \|f(\cdot)\|_{\mathcal{H}(R_1)}, \tag{55}$$

which holds for all $f(\cdot) \in \mathcal{H}(R_1)$.

Let us now return to the MVPs corresponding to $\mathcal{E}$ and $\mathcal{E}'$. From (55) with $\mathcal{D}_1 = \mathcal{X}$, $\mathcal{D}_2 = \mathcal{X}'$, $\mathcal{H}(R_1) = \mathcal{H}_{\mathcal{E},\mathbf{x}_0}$, and $\mathcal{H}(R_2) = \mathcal{H}_{\mathcal{E}',\mathbf{x}_0}$, we can conclude that, for any $\mathbf{x}_0 \in \mathcal{X}'$,

$$\begin{aligned} M'\big(c(\cdot)\big|_{\mathcal{X}'}, \mathbf{x}_0\big) &\overset{(11)}{=} \big\|\gamma(\cdot)\big|_{\mathcal{X}'}\big\|^2_{\mathcal{H}_{\mathcal{E}',\mathbf{x}_0}} - \gamma^2(\mathbf{x}_0) \\ &\overset{(55)}{\leq} \|\gamma(\cdot)\|^2_{\mathcal{H}_{\mathcal{E},\mathbf{x}_0}} - \gamma^2(\mathbf{x}_0) \\ &= M(c(\cdot), \mathbf{x}_0). \end{aligned} \tag{56}$$

Here, we also used the fact that $\gamma(\cdot)\big|_{\mathcal{X}'} = c(\cdot)\big|_{\mathcal{X}'} + g(\cdot)\big|_{\mathcal{X}'}$. The inequality in (56) means that a reduction of the parameter set $\mathcal{X}$ can never result in a deterioration of the achievable performance, i.e., in a higher minimum achievable variance. Besides this rather intuitive fact, the result (53) has the following consequence: Consider an estimation problem $\mathcal{E} = \big(\mathcal{X}, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ whose statistical model $\{f(\mathbf{y};\mathbf{x})\}_{\mathbf{x}\in\mathcal{X}}$ satisfies (8) at some $\mathbf{x}_0 \in \mathcal{X}$ and moreover is contained in a "larger" model $\{f(\mathbf{y};\mathbf{x})\}_{\mathbf{x}\in\tilde{\mathcal{X}}}$ with $\tilde{\mathcal{X}} \supseteq \mathcal{X}$. If the larger model $\{f(\mathbf{y};\mathbf{x})\}_{\mathbf{x}\in\tilde{\mathcal{X}}}$ also satisfies (8), it follows from (53) that a prescribed bias function $c(\cdot) : \mathcal{X} \to \mathbb{R}$ can only be valid for $\mathcal{E}$ at $\mathbf{x}_0$ if it is the restriction of a function $c'(\cdot) : \tilde{\mathcal{X}} \to \mathbb{R}$ that is a valid bias function for the estimation problem $\tilde{\mathcal{E}} = \big(\tilde{\mathcal{X}}, f(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ at $\mathbf{x}_0$. This holds true since every valid bias function for $\mathcal{E}$ at $\mathbf{x}_0$ corresponds to a mean function that is an element of the RKHS $\mathcal{H}_{\mathcal{E},\mathbf{x}_0}$, which by (53) consists precisely of the restrictions of the elements of the RKHS $\mathcal{H}_{\tilde{\mathcal{E}},\mathbf{x}_0}$, which by (10) consists precisely of the mean functions corresponding to bias functions that are valid for $\tilde{\mathcal{E}}$ at $\mathbf{x}_0$.

For the remainder of this section, we restrict our discussion to estimation problems $\mathcal{E}^{(A)} = \big(\mathcal{X}, f^{(A)}(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ whose statistical model is an exponential family model. The next result characterizes the analytic properties of the mean functions $\gamma(\cdot)$ that belong to an RKHS $\mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}$. A proof is provided in Appendix C.

Lemma V.3. *Consider an estimation problem* $\mathcal{E}^{(A)} = \big(\mathcal{X}, f^{(A)}(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ *with an open parameter set* $\mathcal{X} \subseteq \mathcal{N}$ *satisfying* (46) *for some* $\mathbf{x}_0 \in \mathcal{X}$. *Let* $\gamma(\cdot) \in \mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}$ *be such that the partial derivatives* $\frac{\partial^{\mathbf{p}}\gamma(\mathbf{x})}{\partial\mathbf{x}^{\mathbf{p}}}\big|_{\mathbf{x}=\mathbf{x}_0}$ *vanish for every multi-index* $\mathbf{p} \in \mathbb{Z}_+^N$. *Then* $\gamma(\mathbf{x}) = 0$ *for all* $\mathbf{x} \in \mathcal{X}$.

Note that since $\mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}$ is differentiable at $\mathbf{x}_0$ up to any order (see Lemma V.1), it contains the function set $\big\{r^{(\mathbf{p})}_{\mathbf{x}_0}(\mathbf{x})\big\}_{\mathbf{p}\in\mathbb{Z}_+^N}$ defined in (19). Moreover, by (20), for any $f(\cdot) \in \mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}$ and any $\mathbf{p} \in \mathbb{Z}_+^N$, there is $\big\langle r^{(\mathbf{p})}_{\mathbf{x}_0}(\cdot), f(\cdot)\big\rangle_{\mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}} = \frac{\partial^{\mathbf{p}} f(\mathbf{x})}{\partial\mathbf{x}^{\mathbf{p}}}\big|_{\mathbf{x}=\mathbf{x}_0}$. Hence, under the assumptions of Lemma V.3, we have that if a function $f(\cdot) \in \mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}$ satisfies $\big\langle r^{(\mathbf{p})}_{\mathbf{x}_0}(\cdot), f(\cdot)\big\rangle_{\mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}} = 0$ for all $\mathbf{p} \in \mathbb{Z}_+^N$, then $f(\cdot) \equiv 0$. Thus, in this case, the set $\big\{r^{(\mathbf{p})}_{\mathbf{x}_0}(\mathbf{x})\big\}_{\mathbf{p}\in\mathbb{Z}_+^N}$ is complete for the RKHS $\mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}$.

Upon combining (53) and (54) with Lemma V.3, we arrive at the second main result of this section:

Theorem V.4. *Consider an estimation problem* $\mathcal{E}^{(A)} = \big(\mathcal{X}, f^{(A)}(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ *with an open parameter set* $\mathcal{X} \subseteq \mathcal{N}$ *satisfying* (46) *for some* $\mathbf{x}_0 \in \mathcal{X}$*, and a prescribed bias function* $c(\cdot)$ *that is valid for* $\mathcal{E}^{(A)}$ *at* $\mathbf{x}_0$. *Furthermore consider a reduced parameter set* $\mathcal{X}_1 \subseteq \mathcal{X}$ *such that* $\mathbf{x}_0 \in \mathcal{X}_1^{\mathrm{o}}$. *Let* $\mathcal{E}_1^{(A)} \triangleq \big(\mathcal{X}_1, f^{(A)}(\mathbf{y};\mathbf{x}), g(\cdot)\big)$ *denote the estimation problem that is obtained from* $\mathcal{E}^{(A)}$ *by reducing the parameter set to* $\mathcal{X}_1$*, and let* $c_1(\cdot) \triangleq c(\cdot)\big|_{\mathcal{X}_1}$. *Then, the minimum achievable variance for the restricted estimation problem* $\mathcal{E}_1^{(A)}$ *and the restricted bias function* $c_1(\cdot)$*, denoted by* $M_1(c_1(\cdot), \mathbf{x}_0)$*, is equal to the minimum achievable variance for the original estimation problem* $\mathcal{E}^{(A)}$ *and the original bias function* $c(\cdot)$*, i.e.,*

$$M_1(c_1(\cdot), \mathbf{x}_0) = M(c(\cdot), \mathbf{x}_0).$$

A proof of this theorem is provided in Appendix D. Note that the requirement $\mathbf{x}_0 \in \mathcal{X}_1^{\mathrm{o}}$ of the theorem implies that the reduced parameter set $\mathcal{X}_1$ must contain a neighborhood of $\mathbf{x}_0$, i.e., an open ball $\mathcal{B}(\mathbf{x}_0, r)$ with some radius $r > 0$. The message of the theorem is that, for an estimation problem based on an exponential family, parameter set reductions have no effect on the minimum achievable variance at $\mathbf{x}_0$ as long as the reduced parameter set contains a neighborhood of $\mathbf{x}_0$.

VI. Conclusion

The mathematical framework of reproducing kernel Hilbert spaces (RKHS) provides powerful tools for the analysis of minimum variance estimation (MVE) problems. Building upon the theoretical foundation developed in the seminal papers [2] and [3], we derived novel results concerning the RKHS-based analysis of lower variance bounds for MVE, of sufficient statistics, and of MVE problems conforming to an exponential family of distributions. More specifically, we presented an RKHS-based geometric interpretation of several well-known lower bounds on the estimator variance. We showed that each of these bounds is related to the orthogonal projection onto an associated subspace of the RKHS. In particular, the subspace associated with the Cramér–Rao bound is based on the strong structural properties of a *differentiable* RKHS. For a wide class of estimation problems, we proved that the minimum achievable variance, which is the tightest possible lower bound on the estimator variance (Barankin bound), is a lower semi-continuous function of the parameter vector. In some cases, this fact can be used to show that a given lower bound on the estimator variance is not maximally tight. Furthermore, we proved that the RKHS associated with an estimation problem remains unchanged if the observation is replaced by a sufficient statistic.

Finally, we specialized the RKHS description to estimation problems whose observation conforms to an exponential family of distributions. We showed that the kernel of the RKHS has a particularly simple expression in terms of the moment-generating function of the exponential family, and that the RKHS itself is differentiable up to any order. Using this differentiability, we derived novel closed-form lower bounds on the estimator variance. We also showed that a reduction of the parameter set has no effect on the minimum achievable variance at a given reference parameter vector $\mathbf{x}_0$ if the reduced parameter set contains a neighborhood of $\mathbf{x}_0$.

Promising directions for future work include the practical implementation of message passing algorithms for the efficient computation of the lower variance bounds for exponential families derived in Section V-C. Furthermore, in view of the close relations between exponential families and probabilistic graphical models [43], it would be interesting to explore the relations between the graph-theoretic properties of the graph associated with an exponential family and the properties of the RKHS associated with that exponential family.

APPENDIX A
PROOF OF THEOREM III.6

We first note that our assumption that the prescribed bias function $c(\cdot)$ is valid for $\mathcal{E}$ at every $\mathbf{x} \in \mathcal{C}$ has two consequences. First, $M(c(\cdot), \mathbf{x}) < \infty$ for every $\mathbf{x} \in \mathcal{C}$ (cf. our definition of the validity of a bias function in Section II-A); second, due to (10), the prescribed mean function $\gamma(\cdot) = c(\cdot) + g(\cdot)$ belongs to $\mathcal{H}_{\mathcal{E},\mathbf{x}}$ for every $\mathbf{x} \in \mathcal{C}$.

Following [2], we define the *linear span of a kernel function* $R(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, denoted by $\mathcal{L}(R)$, as the set of all functions $f(\cdot) : \mathcal{X} \to \mathbb{R}$ that are finite linear combinations of the form

$$f(\cdot) = \sum_{l\in[L]} a_l R(\cdot, \mathbf{x}_l)\,, \quad \text{with } \mathbf{x}_l \in \mathcal{X},\, a_l \in \mathbb{R},\, L \in \mathbb{N}\,. \tag{57}$$

The linear span $\mathcal{L}(R)$ can be used to express the norm of any function $h(\cdot) \in \mathcal{H}(R)$ according to

$$\|h(\cdot)\|^2_{\mathcal{H}(R)} = \sup_{\substack{f(\cdot)\in\mathcal{L}(R)\\ \|f(\cdot)\|^2_{\mathcal{H}(R)}>0}} \frac{\langle h(\cdot), f(\cdot)\rangle^2_{\mathcal{H}(R)}}{\|f(\cdot)\|^2_{\mathcal{H}(R)}}\,. \tag{58}$$

This expression can be shown by combining [10, Theorem 3.1.2] and [10, Theorem 3.2.2]. We can now develop the minimum achievable variance $M(c(\cdot), \mathbf{x})$ as follows:

$$\begin{aligned} M(c(\cdot), \mathbf{x}) &\overset{(11)}{=} \|\gamma(\cdot)\|^2_{\mathcal{H}_{\mathcal{E},\mathbf{x}}} - \gamma^2(\mathbf{x}) \\ &\overset{(58)}{=} \sup_{\substack{f(\cdot)\in\mathcal{L}(R_{\mathcal{E},\mathbf{x}})\\ \|f(\cdot)\|^2_{\mathcal{H}_{\mathcal{E},\mathbf{x}}}>0}} \frac{\langle \gamma(\cdot), f(\cdot)\rangle^2_{\mathcal{H}_{\mathcal{E},\mathbf{x}}}}{\|f(\cdot)\|^2_{\mathcal{H}_{\mathcal{E},\mathbf{x}}}} - \gamma^2(\mathbf{x})\,. \end{aligned}$$

Using (57) and letting $\mathcal{D} \triangleq \{\mathbf{x}_1, \dots, \mathbf{x}_L\}$, $\mathbf{a} \triangleq (a_1 \cdots a_L)^T$, and $\mathcal{A}_{\mathcal{D}} \triangleq \big\{\mathbf{a} \in \mathbb{R}^L \big| \sum_{l,l'\in[L]} a_l a_{l'} R_{\mathcal{E},\mathbf{x}}(\mathbf{x}_l, \mathbf{x}_{l'}) > 0\big\}$, we obtain further

$$M(c(\cdot), \mathbf{x}) = \sup_{\mathcal{D}\subseteq\mathcal{X},\, L\in\mathbb{N},\, \mathbf{a}\in\mathcal{A}_{\mathcal{D}}} h_{\mathcal{D},\mathbf{a}}(\mathbf{x})\,. \tag{59}$$

Here, our notation $\sup_{\mathcal{D}\subseteq\mathcal{X},\, L\in\mathbb{N},\, \mathbf{a}\in\mathcal{A}_{\mathcal{D}}}$ indicates that the supremum is taken not only with respect to the elements $\mathbf{x}_l$ of $\mathcal{D}$ but also with respect to the size of $\mathcal{D}$, $L = |\mathcal{D}|$; furthermore, the function $h_{\mathcal{D},\mathbf{a}}(\cdot) : \mathcal{X} \to \mathbb{R}$ is given by

$$\begin{aligned} h_{\mathcal{D},\mathbf{a}}(\mathbf{x}) &\triangleq \frac{\big\langle \gamma(\cdot), \sum_{l\in[L]} a_l R_{\mathcal{E},\mathbf{x}}(\cdot, \mathbf{x}_l)\big\rangle^2_{\mathcal{H}_{\mathcal{E},\mathbf{x}}}}{\big\| \sum_{l\in[L]} a_l R_{\mathcal{E},\mathbf{x}}(\cdot, \mathbf{x}_l)\big\|^2_{\mathcal{H}_{\mathcal{E},\mathbf{x}}}} - \gamma^2(\mathbf{x}) \\ &= \frac{\big(\sum_{l\in[L]} a_l \big\langle \gamma(\cdot), R_{\mathcal{E},\mathbf{x}}(\cdot, \mathbf{x}_l)\big\rangle_{\mathcal{H}_{\mathcal{E},\mathbf{x}}}\big)^2}{\sum_{l,l'\in[L]} a_l a_{l'} \big\langle R_{\mathcal{E},\mathbf{x}}(\cdot, \mathbf{x}_l)\, R_{\mathcal{E},\mathbf{x}}(\cdot, \mathbf{x}_{l'})\big\rangle_{\mathcal{H}_{\mathcal{E},\mathbf{x}}}} \\ &\qquad\qquad - \gamma^2(\mathbf{x}) \\ &\overset{(4)}{=} \frac{\big(\sum_{l\in[L]} a_l \gamma(\mathbf{x}_l)\big)^2}{\sum_{l,l'\in[L]} a_l a_{l'} R_{\mathcal{E},\mathbf{x}}(\mathbf{x}_l, \mathbf{x}_{l'})} - \gamma^2(\mathbf{x})\,. \end{aligned}$$

For any finite set $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_L\} \subseteq \mathcal{X}$ and any $\mathbf{a} \in \mathcal{A}_{\mathcal{D}}$, it follows from our assumptions of continuity of $R_{\mathcal{E},\mathbf{x}}(\cdot,\cdot)$ with respect to $\mathbf{x}$ on $\mathcal{C}$ (see (35)) and continuity of $\gamma(\mathbf{x})$ on $\mathcal{C}$ that the function $h_{\mathcal{D},\mathbf{a}}(\mathbf{x})$ is continuous in a neighborhood around any point $\mathbf{x}_0 \in \mathcal{C}$. Thus, for any $\mathbf{x}_0 \in \mathcal{C}$, there exists a radius $\delta_0 > 0$ such that $h_{\mathcal{D},\mathbf{a}}(\mathbf{x})$ is continuous on $\mathcal{B}(\mathbf{x}_0, \delta_0) \subseteq \mathcal{C}$.

We will now show that the function $M(c(\cdot), \mathbf{x})$ given by (59) is lower semi-continuous at every $\mathbf{x}_0 \in \mathcal{C}$, i.e., for any $\mathbf{x}_0 \in \mathcal{C}$ and $\varepsilon > 0$, we can find a radius $r > 0$ such that

$$M(c(\cdot), \mathbf{x}) \geq M(c(\cdot), \mathbf{x}_0) - \varepsilon\,, \quad \text{for all } \mathbf{x} \in \mathcal{B}(\mathbf{x}_0, r)\,. \tag{60}$$

Due to (59), there must be a finite subset $\mathcal{D}_0 \subseteq \mathcal{X}$ and a vector $\mathbf{a}_0 \in \mathcal{A}_{\mathcal{D}_0}$ such that[6]

$$h_{\mathcal{D}_0,\mathbf{a}_0}(\mathbf{x}_0) \geq M(c(\cdot), \mathbf{x}_0) - \frac{\varepsilon}{2}\,, \tag{61}$$

for any given $\varepsilon > 0$. Furthermore, since $h_{\mathcal{D}_0,\mathbf{a}_0}(\mathbf{x})$ is continuous on $\mathcal{B}(\mathbf{x}_0, \delta_0)$ as shown above, there is a radius $r_0 > 0$ (with $r_0 \leq \delta_0$) such that

$$h_{\mathcal{D}_0,\mathbf{a}_0}(\mathbf{x}) \geq h_{\mathcal{D}_0,\mathbf{a}_0}(\mathbf{x}_0) - \frac{\varepsilon}{2}\,, \quad \text{for all } \mathbf{x} \in \mathcal{B}(\mathbf{x}_0, r_0)\,. \tag{62}$$

[6]Indeed, if (61) were not true, we would have $h_{\mathcal{D},\mathbf{a}}(\mathbf{x}_0) < M(c(\cdot), \mathbf{x}_0) - \varepsilon/2$ for every choice of $\mathcal{D}$ and $\mathbf{a}$. This, in turn, would imply that

$$\sup_{\mathcal{D}\subseteq\mathcal{X},\, L\in\mathbb{N},\, \mathbf{a}\in\mathcal{A}_{\mathcal{D}}} h_{\mathcal{D},\mathbf{a}}(\mathbf{x}_0) \leq M(c(\cdot), \mathbf{x}_0) - \frac{\varepsilon}{2} < M(c(\cdot), \mathbf{x}_0)\,,$$

yielding the contradiction

$$M(c(\cdot), \mathbf{x}_0) \overset{(59)}{=} \sup_{\mathcal{D}\subseteq\mathcal{X},\, L\in\mathbb{N},\, \mathbf{a}\in\mathcal{A}_{\mathcal{D}}} h_{\mathcal{D},\mathbf{a}}(\mathbf{x}_0) < M(c(\cdot), \mathbf{x}_0)\,.$$

By combining this inequality with (61), it follows that there is a radius $r>0$ (with $r\leq\delta_0$) such that for any $\mathbf{x}\in\mathcal{B}(\mathbf{x}_0,r)$ we have

$$h_{\mathcal{D}_0,\mathbf{a}_0}(\mathbf{x}) \overset{(62)}{\geq} h_{\mathcal{D}_0,\mathbf{a}_0}(\mathbf{x}_0)-\frac{\varepsilon}{2} \overset{(61)}{\geq} M(c(\cdot),\mathbf{x}_0)-\varepsilon\,, \quad (63)$$

and further

$$\begin{aligned} M(c(\cdot),\mathbf{x}) &\overset{(59)}{=} \sup_{\mathcal{D}\subseteq\mathcal{X},\, L\in\mathbb{N},\, \mathbf{a}\in\mathcal{A}_{\mathcal{D}}} h_{\mathcal{D},\mathbf{a}}(\mathbf{x}) \\ &\geq h_{\mathcal{D}_0,\mathbf{a}_0}(\mathbf{x}) \\ &\overset{(63)}{\geq} M(c(\cdot),\mathbf{x}_0)-\varepsilon\,. \end{aligned}$$

Thus, for any given $\varepsilon>0$, there is a radius $r>0$ (with $r\leq\delta_0$) such that $M(c(\cdot),\mathbf{x})\geq M(c(\cdot),\mathbf{x}_0)-\varepsilon$ for all $\mathbf{x}\in\mathcal{B}(\mathbf{x}_0,r)$, i.e., (60) has been proved.

Appendix B
Proof of Theorem V.2

The bound (48) in Theorem V.2 is derived by using an isometry between the RKHS $\mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}$ and the RKHS $\mathcal{H}(R)$ that is defined by the kernel

$$R(\cdot,\cdot):\mathcal{X}\times\mathcal{X}\to\mathbb{R}\,,\quad R(\mathbf{x}_1,\mathbf{x}_2)=\frac{\lambda(\mathbf{x}_1+\mathbf{x}_2-\mathbf{x}_0)}{\lambda(\mathbf{x}_0)}\,. \quad (64)$$

It is easily verified that $R(\cdot,\cdot)$ and, thus, $\mathcal{H}(R)$ are differentiable up to any order. Invoking [10, Theorem 3.3.4], it can be shown that the two RKHSs $\mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}$ and $\mathcal{H}(R)$ are isometric and a specific congruence $\mathsf{J}:\mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}\to\mathcal{H}(R)$ is given by

$$\mathsf{J}[f(\cdot)]=\frac{\lambda(\mathbf{x})}{\lambda(\mathbf{x}_0)}\,f(\mathbf{x})\,,\quad f(\cdot)\in\mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}\,. \quad (65)$$

Similarly to the bound (21), we can then obtain a lower bound on $v(\hat{g}(\cdot);\mathbf{x}_0)$ via an orthogonal projection onto a subspace of $\mathcal{H}(R)$. Indeed, with $c(\cdot)=\gamma(\cdot)-g(\cdot)$ denoting the bias function of the estimator $\hat{g}(\cdot)$, we have

$$\begin{aligned} v(\hat{g}(\cdot);\mathbf{x}_0) &\overset{(3)}{\geq} M(c(\cdot),\mathbf{x}_0) \\ &\overset{(11)}{=} \|\gamma(\cdot)\|^2_{\mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}}-\gamma^2(\mathbf{x}_0) \\ &\overset{(a)}{=} \big\|\mathsf{J}[\gamma(\cdot)]\big\|^2_{\mathcal{H}(R)}-\gamma^2(\mathbf{x}_0) \\ &\geq \big\|\big(\mathsf{J}[\gamma(\cdot)]\big)_{\mathcal{U}}\big\|^2_{\mathcal{H}(R)}-\gamma^2(\mathbf{x}_0)\,, \end{aligned} \quad (66)$$

for an arbitrary subspace $\mathcal{U}\subseteq\mathcal{H}(R)$. Here, step (a) is due to the fact that J is a congruence, and $(\cdot)_{\mathcal{U}}$ denotes orthogonal projection onto $\mathcal{U}$. The bound (48) is obtained from (66) by choosing the subspace as $\mathcal{U}\triangleq\mathrm{span}\big\{r_{\mathbf{x}_0}^{(\mathbf{p}_l)}(\cdot)\big\}_{l\in[L]}$, with the functions $r_{\mathbf{x}_0}^{(\mathbf{p}_l)}(\cdot)\in\mathcal{H}(R)$ as defined in (19), i.e., $r_{\mathbf{x}_0}^{(\mathbf{p}_l)}(\mathbf{x})=\frac{\partial^{\mathbf{p}_l}R(\mathbf{x},\mathbf{x}_2)}{\partial\mathbf{x}_2^{\mathbf{p}_l}}\Big|_{\mathbf{x}_2=\mathbf{x}_0}$.

Let us denote the image of $\gamma(\cdot)$ under the isometry J by $\tilde{\gamma}(\cdot)\triangleq\mathsf{J}[\gamma(\cdot)]$. According to (65),

$$\tilde{\gamma}(\mathbf{x})=\frac{\lambda(\mathbf{x})}{\lambda(\mathbf{x}_0)}\,\gamma(\mathbf{x})\,. \quad (67)$$

Furthermore, the variance bound (66) reads

$$v(\hat{g}(\cdot);\mathbf{x}_0)\,\geq\,\|\tilde{\gamma}_{\mathcal{U}}(\cdot)\|^2_{\mathcal{H}(R)}-\gamma^2(\mathbf{x}_0)\,.$$

Using (15), we obtain further

$$v(\hat{g}(\cdot);\mathbf{x}_0)\,\geq\,\mathbf{n}^T(\mathbf{x}_0)\,\mathbf{S}^{\dagger}(\mathbf{x}_0)\,\mathbf{n}(\mathbf{x}_0)-\gamma^2(\mathbf{x}_0)\,, \quad (68)$$

where, according to (16), the entries of $\mathbf{n}(\mathbf{x}_0)$ and $\mathbf{S}(\mathbf{x}_0)$ are calculated as follows:

$$\begin{aligned} \big(\mathbf{n}(\mathbf{x}_0)\big)_l &\overset{(16)}{=} \big\langle\tilde{\gamma}(\cdot),r_{\mathbf{x}_0}^{(\mathbf{p}_l)}(\cdot)\big\rangle_{\mathcal{H}(R)} \\ &\overset{(20)}{=} \frac{\partial^{\mathbf{p}_l}\tilde{\gamma}(\mathbf{x})}{\partial\mathbf{x}^{\mathbf{p}_l}}\bigg|_{\mathbf{x}=\mathbf{x}_0} \\ &\overset{(67)}{=} \frac{1}{\lambda(\mathbf{x}_0)}\,\frac{\partial^{\mathbf{p}_l}[\lambda(\mathbf{x})\gamma(\mathbf{x})]}{\partial\mathbf{x}^{\mathbf{p}_l}}\bigg|_{\mathbf{x}=\mathbf{x}_0} \\ &\overset{(a)}{=} \frac{1}{\lambda(\mathbf{x}_0)}\sum_{\mathbf{p}\leq\mathbf{p}_l}\binom{\mathbf{p}_l}{\mathbf{p}}\frac{\partial^{\mathbf{p}_l-\mathbf{p}}\lambda(\mathbf{x})}{\partial\mathbf{x}^{\mathbf{p}_l-\mathbf{p}}}\,\frac{\partial^{\mathbf{p}}\gamma(\mathbf{x})}{\partial\mathbf{x}^{\mathbf{p}}}\bigg|_{\mathbf{x}=\mathbf{x}_0} \\ &\overset{(43)}{=} \sum_{\mathbf{p}\leq\mathbf{p}_l}\binom{\mathbf{p}_l}{\mathbf{p}}\,\mathsf{E}_{\mathbf{x}_0}\big\{\boldsymbol{\phi}^{\mathbf{p}_l-\mathbf{p}}(\mathbf{y})\big\}\,\frac{\partial^{\mathbf{p}}\gamma(\mathbf{x})}{\partial\mathbf{x}^{\mathbf{p}}}\bigg|_{\mathbf{x}=\mathbf{x}_0} \end{aligned} \quad (69)$$

(here, (a) is due to the generalized Leibniz rule for differentiation of a product of two functions [13, p. 104]), and

$$\begin{aligned} \big(\mathbf{S}(\mathbf{x}_0)\big)_{l,l'} &\overset{(16)}{=} \big\langle r_{\mathbf{x}_0}^{(\mathbf{p}_l)}(\cdot),r_{\mathbf{x}_0}^{(\mathbf{p}_{l'})}(\cdot)\big\rangle_{\mathcal{H}(R)} \\ &\overset{(20)}{=} \frac{\partial^{\mathbf{p}_l}r_{\mathbf{x}_0}^{(\mathbf{p}_{l'})}(\mathbf{x})}{\partial\mathbf{x}^{\mathbf{p}_l}}\bigg|_{\mathbf{x}=\mathbf{x}_0} \\ &\overset{(19)}{=} \frac{\partial^{\mathbf{p}_l}}{\partial\mathbf{x}^{\mathbf{p}_l}}\bigg\{\frac{\partial^{\mathbf{p}_{l'}}R(\mathbf{x},\mathbf{x}_2)}{\partial\mathbf{x}_2^{\mathbf{p}_{l'}}}\bigg|_{\mathbf{x}_2=\mathbf{x}_0}\bigg\}\bigg|_{\mathbf{x}=\mathbf{x}_0} \\ &\overset{(64)}{=} \frac{1}{\lambda(\mathbf{x}_0)}\,\frac{\partial^{\mathbf{p}_l+\mathbf{p}_{l'}}\lambda(\mathbf{x})}{\partial\mathbf{x}^{\mathbf{p}_l+\mathbf{p}_{l'}}}\bigg|_{\mathbf{x}=\mathbf{x}_0} \\ &\overset{(43)}{=} \mathsf{E}_{\mathbf{x}_0}\big\{\boldsymbol{\phi}^{\mathbf{p}_l+\mathbf{p}_{l'}}(\mathbf{y})\big\}\,. \end{aligned} \quad (70)$$

Note that the application of (20) was based on the differentiability of $\mathcal{H}(R)$. Comparing (68), (69), and (70) with (48), (49), and (50), respectively, we conclude that the theorem is proved.

Appendix C
Proof of Lemma V.3

For $\mathcal{E}^{(A)}=\big(\mathcal{X},f^{(A)}(\mathbf{y};\mathbf{x}),g(\cdot)\big)$ and $\mathbf{x}_0\in\mathcal{X}$, we consider a function $\gamma(\cdot):\mathcal{X}\to\mathbb{R}$ belonging to the RKHS $\mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}$. The idea of the proof is to reduce the characterization of $\gamma(\cdot)$ to that of the LMV estimator $\hat{g}^{(\mathbf{x}_0)}(\cdot)$ for prescribed mean function $\gamma(\cdot)$. This estimator is guaranteed to exist since $\gamma(\cdot)\in\mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}$, and it is given by $\hat{g}^{(\mathbf{x}_0)}(\cdot)=\mathsf{J}[\gamma(\cdot)]$ (see (12)). Trivially, $v\big(\hat{g}^{(\mathbf{x}_0)}(\cdot);\mathbf{x}_0\big)=M(c(\cdot),\mathbf{x}_0)<\infty$ and $\mathsf{E}_{\mathbf{x}}\big\{\hat{g}^{(\mathbf{x}_0)}(\mathbf{y})\big\}=\gamma(\mathbf{x})$ for all $\mathbf{x}\in\mathcal{X}$. For later use, we also note that

$$\begin{aligned} \mathsf{E}_{\mathbf{x}_0}\big\{\big(\hat{g}^{(\mathbf{x}_0)}(\mathbf{y})\big)^2\big\} &= v\big(\hat{g}^{(\mathbf{x}_0)}(\mathbf{y});\mathbf{x}_0\big)+\big(\mathsf{E}_{\mathbf{x}_0}\big\{\hat{g}^{(\mathbf{x}_0)}(\mathbf{y})\big\}\big)^2 \\ &= M(c(\cdot),\mathbf{x}_0)+\gamma^2(\mathbf{x}_0) \\ &< \infty\,. \end{aligned} \quad (71)$$

For any exponential family based estimation problem $\mathcal{E}^{(A)}=\big(\mathcal{X},f^{(A)}(\mathbf{y};\mathbf{x}),g(\cdot)\big)$, it follows from [42, Theorem 2.7] that the mean function $\mathsf{E}_{\mathbf{x}}\{\hat{g}(\cdot)\}$ of any estimator $\hat{g}(\cdot)$

is analytic[7] on the interior $\mathcal{T}^{\text{o}}$ of the set $\mathcal{T} \triangleq \big\{\mathbf{x} \in \mathcal{N} \big| \mathsf{E}_{\mathbf{x}}\{|\hat{g}(\mathbf{y})|\} < \infty\big\}$. Furthermore, $\mathcal{T}$ can be shown to be a convex set [42, Corollary 2.6]. In particular, the mean function $\gamma(\mathbf{x})$ of the LMV estimator $\hat{g}^{(\mathbf{x}_0)}(\cdot)$ is analytic on the interior $\mathcal{T}_0^{\text{o}}$ of the convex set $\mathcal{T}_0 \triangleq \big\{\mathbf{x} \in \mathcal{N} \big| \mathsf{E}_{\mathbf{x}}\{|\hat{g}^{(\mathbf{x}_0)}(\mathbf{y})|\} < \infty\big\}$. However, here we are interested in the behavior of $\mathsf{E}_{\mathbf{x}}\big\{\hat{g}^{(\mathbf{x}_0)}(\mathbf{y})\big\} = \gamma(\mathbf{x})$ on the domain $\mathcal{X}$. Therefore, we will next verify that $\mathcal{X} \subseteq \mathcal{T}_0$.

Using the Hilbert space $\mathcal{H}_{\mathbf{x}_0} \triangleq \big\{t(\mathbf{y}) \big| \mathsf{E}_{\mathbf{x}_0}\{t^2(\mathbf{y})\} < \infty\big\}$ and associated inner product $\langle t_1(\mathbf{y}), t_2(\mathbf{y})\rangle_{\text{RV}} = \mathsf{E}_{\mathbf{x}_0}\{t_1(\mathbf{y})t_2(\mathbf{y})\}$, we obtain for an arbitrary $\mathbf{x} \in \mathcal{X} \subseteq \mathcal{N}$

$$
\begin{aligned}
&\mathsf{E}_{\mathbf{x}}\big\{\big|\hat{g}^{(\mathbf{x}_0)}(\mathbf{y})\big|\big\} \\
&= \mathsf{E}_{\mathbf{x}_0}\bigg\{\big|\hat{g}^{(\mathbf{x}_0)}(\mathbf{y})\big| \, \frac{f^{(A)}(\mathbf{y};\mathbf{x})}{f^{(A)}(\mathbf{y};\mathbf{x}_0)}\bigg\} \\
&= \big\langle \big|\hat{g}^{(\mathbf{x}_0)}(\mathbf{y})\big|, \rho_{\mathcal{E}^{(A)},\mathbf{x}_0}(\mathbf{y},\mathbf{x})\big\rangle_{\text{RV}} \\
&\overset{(a)}{\leq} \sqrt{\big\langle \big|\hat{g}^{(\mathbf{x}_0)}(\mathbf{y})\big|, \big|\hat{g}^{(\mathbf{x}_0)}(\mathbf{y})\big|\big\rangle_{\text{RV}} \big\langle \rho_{\mathcal{E}^{(A)},\mathbf{x}_0}(\mathbf{y},\mathbf{x}), \rho_{\mathcal{E}^{(A)},\mathbf{x}_0}(\mathbf{y},\mathbf{x})\big\rangle_{\text{RV}}} \\
&= \sqrt{\mathsf{E}_{\mathbf{x}_0}\big\{\big(\hat{g}^{(\mathbf{x}_0)}(\mathbf{y})\big)^2\big\} \, \big\langle \rho_{\mathcal{E}^{(A)},\mathbf{x}_0}(\mathbf{y},\mathbf{x}), \rho_{\mathcal{E}^{(A)},\mathbf{x}_0}(\mathbf{y},\mathbf{x})\big\rangle_{\text{RV}}} \\
&\overset{(71),(8)}{<} \infty \,,
\end{aligned}
$$

where (a) follows from the Cauchy-Schwarz inequality in the Hilbert space $\mathcal{H}_{\mathbf{x}_0}$. Thus, we have verified that $\mathsf{E}_{\mathbf{x}}\big\{\big|\hat{g}^{(\mathbf{x}_0)}(\mathbf{y})\big|\big\} < \infty$ on $\mathcal{X}$, which implies that $\mathcal{X} \subseteq \mathcal{T}_0$. Moreover, we have

$$
\mathcal{X} \subseteq \mathcal{T}_0^{\text{o}}. \tag{72}
$$

This is implied[8] by $\mathcal{X} \subseteq \mathcal{T}_0$ together with the fact that (by assumption) $\mathcal{X}$ is an open set.

Next, we will show that because of the analyticity of $\gamma(\cdot)$ on $\mathcal{T}_0^{\text{o}}$, the fact that all partial derivatives of $\gamma(\cdot)$ vanish at $\mathbf{x}_0$ implies that $\gamma(\mathbf{x})=0$ for all $\mathbf{x} \in \mathcal{T}_0^{\text{o}}$, and, in turn, by (72) that $\gamma(\mathbf{x})=0$ for all $\mathbf{x} \in \mathcal{X}$. Let us consider the restrictions

$$
\gamma_{\mathcal{R}_{\mathbf{x}_1}}(a) \triangleq \gamma\big(a\mathbf{x}_1 + (1-a)\mathbf{x}_0\big)\,, \quad a \in (-\varepsilon, 1+\varepsilon)\,, \tag{73}
$$

of $\gamma(\cdot)$ on line segments $\mathcal{R}_{\mathbf{x}_1} \triangleq \big\{a\mathbf{x}_1 + (1-a)\mathbf{x}_0 \big| a \in (-\varepsilon, 1+\varepsilon)\big\}$, where $\mathbf{x}_1 \in \mathcal{T}_0^{\text{o}}$ and $\varepsilon > 0$. Here, ε is chosen sufficiently small such that the vectors $\mathbf{x}_a \triangleq \mathbf{x}_0 - \varepsilon(\mathbf{x}_1 - \mathbf{x}_0)$ and $\mathbf{x}_b \triangleq \mathbf{x}_1 + \varepsilon(\mathbf{x}_1 - \mathbf{x}_0)$ belong to $\mathcal{T}_0^{\text{o}}$, i.e., $\mathbf{x}_a, \mathbf{x}_b \in \mathcal{T}_0^{\text{o}}$. Such an ε can always be found, since $\mathbf{x}_0 \in \mathcal{T}_0^{\text{o}}$ (due to (72)) and $\mathbf{x}_1 \in \mathcal{T}_0^{\text{o}}$ (by assumption). As can be verified easily, any vector in $\mathcal{R}_{\mathbf{x}_1}$ is a convex combination of the vectors $\mathbf{x}_a$ and $\mathbf{x}_b$, which both belong to $\mathcal{T}_0^{\text{o}}$. Therefore we have $\mathcal{R}_{\mathbf{x}_1} \subseteq \mathcal{T}_0^{\text{o}}$ for any $\mathbf{x}_1 \in \mathcal{T}_0^{\text{o}}$, as the interior $\mathcal{T}_0^{\text{o}}$ of the convex set $\mathcal{T}_0$ is itself a convex set [44, Theorem 6.2],[9] i.e., the interior $\mathcal{T}_0^{\text{o}}$ contains any convex combination of its elements.

The function $\gamma_{\mathcal{R}_{\mathbf{x}_1}}(\cdot)\colon (-\varepsilon, 1+\varepsilon) \to \mathbb{R}$ in (73) is the composition of the mean function $\gamma(\cdot)\colon \mathcal{X} \to \mathbb{R}$, which is analytic on $\mathcal{T}_0^{\text{o}}$, with the vector-valued function $\mathbf{b}(\cdot)\colon (-\varepsilon, 1+\varepsilon) \to \mathcal{T}_0^{\text{o}}$ given by $\mathbf{b}(a) = a\mathbf{x}_1 + (1-a)\mathbf{x}_0$. Since each component $b_l(\cdot)$ of the function $\mathbf{b}(\cdot)$, whose domain is the open interval $(-\varepsilon, 1+\varepsilon)$, is an analytic function, the function $\gamma_{\mathcal{R}_{\mathbf{x}_1}}(\cdot)$ is itself analytic [14, Proposition 2.2.8].

Since the partial derivatives of $\gamma(\cdot)$ at $\mathbf{x}_0$, $\frac{\partial^{\mathbf{p}}\gamma(\mathbf{x})}{\partial \mathbf{x}^{\mathbf{p}}}\big|_{\mathbf{x}=\mathbf{x}_0}$, are assumed to vanish for every $\mathbf{p} \in \mathbb{Z}_+^N$, the (ordinary) derivatives of arbitrary order of the scalar function $\gamma_{\mathcal{R}_{\mathbf{x}_1}}(a)$ vanish at $a = 0$ (cf. [13, Theorem 9.15]). According to [14, Corollary 1.2.5], since $\gamma_{\mathcal{R}_{\mathbf{x}_1}}(a)$ is an analytic function, this implies that $\gamma_{\mathcal{R}_{\mathbf{x}_1}}(a)$ vanishes everywhere on its open domain $(-\varepsilon, 1+\varepsilon)$. This, in turn, implies that $\gamma(\cdot)$ vanishes on every line segment $\mathcal{R}_{\mathbf{x}_1}$ with arbitrary $\mathbf{x}_1 \in \mathcal{T}_0^{\text{o}}$ and, thus, $\gamma(\cdot)$ vanishes everywhere on $\mathcal{T}_0^{\text{o}}$. By (72), we finally conclude that $\gamma(\cdot)$ vanishes everywhere on $\mathcal{X}$.

Appendix D
Proof of Theorem V.4

Because $c(\cdot)$ was assumed valid at $\mathbf{x}_0$, the corresponding mean function $\gamma(\cdot) = c(\cdot) + g(\cdot)$ is an element of $\mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}$ (see (10)). Let $\gamma_1(\cdot) \triangleq \gamma(\cdot)\big|_{\mathcal{X}_1}$, and note that $\gamma_1(\cdot)$ is the mean function corresponding to the restricted bias function $c_1(\cdot)$, i.e., $\gamma_1(\cdot) = c_1(\cdot) + g(\cdot)\big|_{\mathcal{X}_1}$. We have $\gamma_1(\cdot) \in \mathcal{H}_{\mathcal{E}_1^{(A)},\mathbf{x}_0}$ due to (10), because $\gamma_1(\mathbf{x})$ is the mean function (evaluated for $\mathbf{x} \in \mathcal{X}_1$) of an estimator $\hat{g}(\cdot)$ that has finite variance at $\mathbf{x}_0 \in \mathcal{X}_1^{\text{o}}$ and whose bias function on $\mathcal{X}$ equals $c(\mathbf{x})$. (The existence of such an estimator $\hat{g}(\cdot)$ is guaranteed since $c(\cdot)$ was assumed valid at $\mathbf{x}_0$.) For the minimum achievable variance for the restricted estimation problem, we obtain

$$
\begin{aligned}
M_1(c_1(\cdot), \mathbf{x}_0) &\overset{(11)}{=} \|\gamma_1(\cdot)\|^2_{\mathcal{H}_{\mathcal{E}_1^{(A)},\mathbf{x}_0}} - \gamma_1^2(\mathbf{x}_0) \\
&\overset{(54)}{=} \min_{\substack{\gamma'(\cdot) \in \mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0} \\ \gamma'(\cdot)|_{\mathcal{X}_1} = \gamma_1(\cdot)}} \|\gamma'(\cdot)\|^2_{\mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}} - \gamma_1^2(\mathbf{x}_0)\,.
\end{aligned} \tag{74}
$$

However, the only function $\gamma'(\cdot) \in \mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}$ that satisfies $\gamma'(\cdot)\big|_{\mathcal{X}_1} = \gamma_1(\cdot)$ is the mean function $\gamma(\cdot)$. This is a consequence of Lemma V.3 and can be verified as follows. Consider a function $\gamma'(\cdot) \in \mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}$ that satisfies $\gamma'(\cdot)\big|_{\mathcal{X}_1} = \gamma_1(\cdot)$. By the definition of $\gamma_1(\cdot)$, we also have $\gamma(\cdot)\big|_{\mathcal{X}_1} = \gamma_1(\cdot)$. Therefore, the difference $\gamma''(\cdot) \triangleq \gamma'(\cdot) - \gamma(\cdot) \in \mathcal{H}_{\mathcal{E}^{(A)},\mathbf{x}_0}$ satisfies $\gamma''(\cdot)\big|_{\mathcal{X}_1} = \gamma'(\cdot)\big|_{\mathcal{X}_1} - \gamma(\cdot)\big|_{\mathcal{X}_1} = \gamma_1(\cdot) - \gamma_1(\cdot) = 0$, i.e., $\gamma''(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{X}_1$. Since $\mathbf{x}_0 \in \mathcal{X}_1^{\text{o}}$, this implies that $\frac{\partial^{\mathbf{p}}\gamma''(\mathbf{x})}{\partial \mathbf{x}^{\mathbf{p}}}\big|_{\mathbf{x}=\mathbf{x}_0} = 0$ for all $\mathbf{p} \in \mathbb{Z}_+^N$. It then follows from Lemma V.3 that $\gamma''(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{X}$ and, thus, $\gamma'(\mathbf{x}) = \gamma(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. This shows that $\gamma(\cdot)$ is the unique

[7]Following [14, Definition 2.2.1], we call a function $f(\cdot)\colon \mathcal{U} \to \mathbb{R}$ defined on some open domain $\mathcal{U} \subseteq \mathbb{R}^N$ *analytic* if for every point $\mathbf{x}_c \in \mathcal{U}$ there exists a power series $\sum_{\mathbf{p} \in \mathbb{Z}_+^N} a_{\mathbf{p}}(\mathbf{x} - \mathbf{x}_c)^{\mathbf{p}}$ converging to $f(\mathbf{x})$ for every $\mathbf{x}$ in some neighborhood of $\mathbf{x}_c$. Note that the coefficients $a_{\mathbf{p}}$ may depend on $\mathbf{x}_c$.

[8]Indeed, assume that the open set $\mathcal{X} \subseteq \mathcal{T}_0$ contains a vector $\mathbf{x}' \in \mathcal{X}$ that does not belong to the interior $\mathcal{T}_0^{\text{o}}$. It follows that no single neighborhood of $\mathbf{x}'$ can be contained in $\mathcal{T}_0$ and, thus, no single neighborhood of $\mathbf{x}'$ can be contained in $\mathcal{X}$, since $\mathcal{X} \subseteq \mathcal{T}_0$. However, because $\mathbf{x}'$ belongs to the open set $\mathcal{X} = \mathcal{X}^{\text{o}}$, there must be at least one neighborhood of $\mathbf{x}'$ that is contained in $\mathcal{X}$. Thus, we arrived at a contradiction, which implies that every vector $\mathbf{x}' \in \mathcal{X}$ must belong to $\mathcal{T}_0^{\text{o}}$, or, equivalently, that $\mathcal{X} \subseteq \mathcal{T}_0^{\text{o}}$.

[9]Strictly speaking, [44, Theorem 6.2] states that the *relative interior* of a convex set is a convex set. However, since we assume that $\mathcal{X}$ is open with nonempty interior and therefore, by (72), also $\mathcal{T}_0$ has a nonempty interior, the relative interior of $\mathcal{T}_0$ coincides with the interior of $\mathcal{T}_0$.

function satisfying $\gamma(\cdot)\big|_{\mathcal{X}_1} = \gamma_1(\cdot)$. Therefore, we have

$$\min_{\substack{\gamma'(\cdot)\in\mathcal{H}_{\mathcal{E}(A),\mathbf{x}_0}\\ \gamma'(\cdot)\big|_{\mathcal{X}_1}=\gamma_1(\cdot)}} \|\gamma'(\cdot)\|^2_{\mathcal{H}_{\mathcal{E}(A),\mathbf{x}_0}} = \|\gamma(\cdot)\|^2_{\mathcal{H}_{\mathcal{E}(A),\mathbf{x}_0}},$$

and thus (74) becomes

$$\begin{aligned} M_1(c_1(\cdot),\mathbf{x}_0) &= \|\gamma(\cdot)\|^2_{\mathcal{H}_{\mathcal{E}(A),\mathbf{x}_0}} - \gamma_1^2(\mathbf{x}_0) \\ &= \|\gamma(\cdot)\|^2_{\mathcal{H}_{\mathcal{E}(A),\mathbf{x}_0}} - \gamma^2(\mathbf{x}_0) \\ &\overset{(11)}{=} M(c(\cdot),\mathbf{x}_0)\,. \end{aligned}$$

Here, the second equality is due to the fact that $\gamma_1(\mathbf{x}_0) = \gamma(\mathbf{x}_0)$ (because $\mathbf{x}_0 \in \mathcal{X}_1^{\mathrm{o}}$).

References

[1] P. Billingsley, *Probability and Measure*, 3rd ed. New York: Wiley, 1995.

[2] E. Parzen, "Statistical inference on time series by Hilbert space methods, I." Appl. Math. Stat. Lab., Stanford University, Stanford, CA, Tech. Rep. 23, Jan. 1959.

[3] D. D. Duttweiler and T. Kailath, "RKHS approach to detection and estimation problems – Part V: Parameter estimation," *IEEE Trans. Inf. Theory*, vol. 19, no. 1, pp. 29–37, Jan. 1973.

[4] S. Smale and D.-X. Zhou, "Learning theory estimates via integral operators and their approximations," *Constr. Approx.*, vol. 26, pp. 153–172, 2007.

[5] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bull. Amer. Math. Soc.*, vol. 39, pp. 1–49, 2002.

[6] S. Schmutzhard, A. Jung, F. Hlawatsch, Z. Ben-Haim, and Y. C. Eldar, "A lower bound on the estimator variance for the sparse linear model," in *Proc. 44th Asilomar Conf. Signals, Systems, Computers*, Pacific Grove, CA, Nov. 2010, pp. 1976–1980.

[7] S. Schmutzhard, A. Jung, and F. Hlawatsch, "Minimum variance estimation for the sparse signal in noise model," in *Proc. IEEE ISIT 2011*, St. Petersburg, Russia, Jul.–Aug. 2011, pp. 124–128.

[8] A. Jung, S. Schmutzhard, F. Hlawatsch, and A. O. Hero III, "Performance bounds for sparse parametric covariance estimation in Gaussian models," in *Proc. IEEE ICASSP 2011*, Prague, Czech Republic, May 2011, pp. 4156–4159.

[9] T. Kailath, "RKHS approach to detection and estimation problems – Part I: Deterministic signals in Gaussian noise," *IEEE Trans. Inf. Theory*, vol. 17, no. 5, pp. 530–549, Sep. 1971.

[10] A. Jung, "An RKHS Approach to Estimation with Sparsity Constraints," Ph.D. dissertation, Vienna University of Technology, 2011, available online: arXiv:1311.5768.

[11] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: Johns Hopkins University Press, 1996.

[12] B. R. Gelbaum and J. M. Olmsted, *Counterexamples in Analysis*. Mineola, NY: Dover Publications, 2003.

[13] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York: McGraw-Hill, 1976.

[14] S. G. Krantz and H. R. Parks, *A Primer of Real Analytic Functions*, 2nd ed. Boston, MA: Birkhäuser, 2002.

[15] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York: Springer, 1998.

[16] P. R. Halmos and L. J. Savage, "Application of the Radon-Nikodym theorem to the theory of sufficient statistics," *Ann. Math. Statist.*, vol. 20, no. 2, pp. 225–241, 1949.

[17] I. A. Ibragimov and R. Z. Has'minskii, *Statistical Estimation. Asymptotic Theory.* New York: Springer, 1981.

[18] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[19] Y. C. Eldar, *Rethinking Biased Estimation: Improving Maximum Likelihood and the Cramér–Rao Bound*, ser. Foundations and Trends in Signal Processing. Hanover, MA: Now Publishers, 2007, vol. 1, no. 4.

[20] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. Pacific Grove, CA: Duxbury, 2002.

[21] N. Aronszajn, "Theory of reproducing kernels," *Trans. Am. Math. Soc.*, vol. 68, no. 3, pp. 337–404, May 1950.

[22] W. Rudin, *Real and Complex Analysis*, 3rd ed. New York: McGraw-Hill, 1987.

[23] E. W. Barankin, "Locally best unbiased estimates," *Ann. Math. Statist.*, vol. 20, no. 4, pp. 477–501, 1949.

[24] C. Stein, "Unbiased estimates with minimum variance," *Ann. Math. Statist.*, vol. 21, no. 3, pp. 406–415, 1950.

[25] P. R. Halmos, *Measure Theory*. New York: Springer, 1974.

[26] H.-W. Sun and D.-X. Zhou, "Reproducing kernel Hilbert spaces associated with analytic translation-invariant Mercer kernels," *J. Fourier Anal. Appl.*, vol. 14, no. 1, pp. 89–101, Feb. 2008.

[27] D.-X. Zhou, "Derivative reproducing properties for kernel methods in learning theory," *J. Comput. Appl. Math.*, vol. 220, no. 1–2, pp. 456–463, Oct. 2008.

[28] ——, "Capacity of reproducing kernel spaces in learning theory," *IEEE Trans. Inf. Theory*, vol. 49, pp. 1743–1752, 2003.

[29] R. McAulay and E. Hofstetter, "Barankin bounds on parameter estimation," *IEEE Trans. Inf. Theory*, vol. 17, no. 6, pp. 669–676, Nov. 1971.

[30] H. Cramér, "A contribution to the theory of statistical estimation," *Skand. Akt. Tidskr.*, vol. 29, pp. 85–94, 1946.

[31] C. R. Rao, "Information and the accuracy attainable in the estimation of statistical parameters," *Bull. Calcutta Math. Soc.*, vol. 37, pp. 81–91, 1945.

[32] P. Stoica and B. C. Ng, "On the Cramér–Rao bound under parametric constraints," *IEEE Signal Processing Letters*, vol. 5, no. 7, pp. 177–179, Jul. 1998.

[33] Z. Ben-Haim and Y. Eldar, "On the constrained Cramér–Rao bound with a singular Fisher information matrix," *IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 453–456, June 2009.

[34] T. J. Moore, "A Theory of Cramér–Rao Bounds for Constrained Parametric Models," Ph.D. dissertation, University of Maryland, 2010.

[35] J. S. Abel, "A bound on mean-square-estimate error," *IEEE Trans. Inf. Theory*, vol. 39, no. 5, pp. 1675–1680, Sep. 1993.

[36] A. Bhattacharyya, "On some analogues of the amount of information and their use in statistical estimation," *Shankya: The Indian Journal of Statistics (1933-1960)*, vol. 8, no. 1, pp. 1–14, Nov. 1946.

[37] J. D. Gorman and A. O. Hero, "Lower bounds for parametric estimation with constraints," *IEEE Trans. Inf. Theory*, vol. 36, no. 6, pp. 1285–1301, Nov. 1990.

[38] D. G. Chapman and H. Robbins, "Minimum variance estimation without regularity assumptions," *Ann. Math. Statist.*, vol. 22, no. 4, pp. 581–586, Dec. 1951.

[39] J. M. Hammersley, "On estimating restricted parameters," *J. Roy. Statist. Soc. B*, vol. 12, no. 2, pp. 192–240, 1950.

[40] Z. Ben-Haim and Y. C. Eldar, "The Cramér–Rao bound for estimating a sparse parameter vector," *IEEE Trans. Signal Processing*, vol. 58, pp. 3384–3389, June 2010.

[41] S. Kullback, *Information Theory and Statistics*. Mineola, NY: Dover Publications, 1968.

[42] L. D. Brown, *Fundamentals of Statistical Exponential Families*. Hayward, CA: Institute of Mathematical Statistics, 1986.

[43] M. J. Wainwright and M. I. Jordan, *Graphical Models, Exponential Families, and Variational Inference*, ser. Foundations and Trends in Machine Learning. Hanover, MA: Now Publishers, 2008, vol. 1, no. 1–2.

[44] R. T. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton Univ. Press, 1970.

Alexander Jung (S'11–M'14) received the Diplom-Ingenieur and Dr. techn. degrees in electrical engineering/signal processing from Vienna University of Technology, Vienna, Austria, in 2008 and 2011, respectively. Since 2008, he has been a Research Assistant with the Institute of Telecommunications, Vienna University of Technology. His research interests are in statistical signal processing with emphasis on sparse estimation problems. He received several national awards and a Best Student Paper Award at IEEE ICASSP 2011.

Sebastian Schmutzhard received the Diplom-Ingenieur degree in technical mathematics from the University of Innsbruck, Austria in 2008, and the Dr. rer. nat. degree in mathematics from the University of Vienna, Austria in 2012. His research interests include numerical analysis, special functions of mathematical physics, and signal processing. He is coauthor of a paper that won a Best Student

Paper Award at IEEE ICASSP 2011.

Franz Hlawatsch (S'85–M'88–SM'00–F'12) received the Diplom-Ingenieur, Dr. techn., and Univ.-Dozent (habilitation) degrees in electrical engineering/signal processing from Vienna University of Technology, Vienna, Austria in 1983, 1988, and 1996, respectively. Since 1983, he has been with the Institute of Telecommunications, Vienna University of Technology, where he is currently an Associate Professor. During 1991–1992, as a recipient of an Erwin Schrödinger Fellowship, he spent a sabbatical year with the Department of Electrical Engineering, University of Rhode Island, Kingston, RI, USA. In 1999, 2000, and 2001, he held one-month Visiting Professor positions with INP/ENSEEIHT, Toulouse, France and IRCCyN, Nantes, France. He (co)authored a book, three review papers that appeared in the IEEE SIGNAL PROCESSING MAGAZINE, about 200 refereed scientific papers and book chapters, and three patents. He coedited three books. His research interests include signal processing for sensor networks and wireless communications, statistical signal processing, and compressive signal processing.

Prof. Hlawatsch was Technical Program Co-Chair of EUSIPCO 2004 and served on the technical committees of numerous IEEE conferences. He was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2003 to 2007 and for the IEEE TRANSACTIONS ON INFORMATION THEORY from 2008 to 2011. From 2004 to 2009, he was a member of the IEEE SPCOM Technical Committee. He is coauthor of papers that won an IEEE Signal Processing Society Young Author Best Paper Award and a Best Student Paper Award at IEEE ICASSP 2011.