IEICE TRANS. INF. & SYST., VOL.E105–D, NO.1 JANUARY 2022

PAPER

Kernel-Based Regressors Equivalent to Stochastic Affine Estimators

Akira TANAKA[†a)], **Masanari NAKAMURA**[†b)], *and* **Hideyuki IMAI**[†c)], *Members*

SUMMARY The solution of the ordinary kernel ridge regression, based on the squared loss function and the squared norm-based regularizer, can be easily interpreted as a stochastic linear estimator by considering the autocorrelation prior for an unknown true function. As is well known, a stochastic affine estimator is one of the simplest extensions of the stochastic linear estimator. However, its corresponding kernel regression problem is not revealed so far. In this paper, we give a formulation of the kernel regression problem, whose solution is reduced to a stochastic affine estimator, and also give interpretations of the formulation.

key words: *kernel regression, autocorrelation prior, linear estimators, affine estimators, optimization criterion*

1. Introduction

The kernel ridge regression (KRR) [1]–[3] is still one of useful function estimators in the field of machine learning. The ordinary KRR is defined as the minimizer of the squared loss function for a given training data set and the squared norm-based regularizer defined in a certain reproducing kernel Hilbert space [4] to which an unknown true function is assumed to belong. This regularizer involves two hyperparameters, that specify a model. One is a reproducing kernel and the other is a regularization parameter. Selection of a model, that achieves good generalization performance, is one of crucial topics in this field. As shown in [5], assuming the existence of the autocorrelation function of an unknown true function, the optimal model of the KRR, in terms of the expected squared error, is specified by the autocorrelation function itself (as a kernel) and the variance of additive noise (as a regularization parameter), which immediately leads the equivalence of the KRR and a stochastic linear estimator [6]. Note that the KRR with this optimal model also agrees with the solution of the Gaussian process regression (GPR) [7]–[10] with the zero-mean assumption, which is formulated by an another approach, that is, the conditional expectation based on a given training data set. It is widely believed that the superiority of the GPR against the KRR is that the GPR can identify the variance of estimates. However, almost the same result was also obtained by the theory of reproducing kernel Hilbert spaces only without the assumptions of probabilistic structures as shown in [11]. Therefore, the essential difference between the KRR and the GPR with the zero-mean assumption is their formulations only.

Needless to say, a stochastic affine estimator [6] is one of the simplest extensions of a stochastic linear estimator. In the framework of the GPR, the solution, corresponding to a stochastic affine estimator, is straightforwardly obtained by centering [10], that is, subtraction of mean, as the same with the general framework of a stochastic estimation [6]. On the other hand, the formulation of kernel regression problems, corresponding to a stochastic affine estimator, is not revealed so far. In this paper, we reveal the formulation of the kernel regression problem whose solution is reduced to that of a stochastic affine estimator. Since the formulation of kernel-based regression is quite different from those of stochastic estimators and the GPR, it is expected that our result can provide a novel aspect for further extensions of the kernel regression problems, while the GPR has limited room for extension due to its principle.

2. Overview of Kernel Ridge Regression

In this section, we give an overview of the KRR [1].

Firstly, we briefly review the theory of reproducing kernel Hilbert spaces [4], [12], [13].

Definition 1: [4] Let $\mathbf{R}^d$ be a d-dimensional real vector space and let $\mathcal{H}$ be a class of functions defined on a certain domain $\mathcal{D} \subset \mathbf{R}^d$, forming a Hilbert space of real-valued functions. The function $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$, $(\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathcal{D})$ is called a reproducing kernel of $\mathcal{H}$, if the following two conditions hold.

$$1.\quad \forall \tilde{\boldsymbol{x}} \in \mathcal{D},\quad K(\cdot, \tilde{\boldsymbol{x}}) \in \mathcal{H} \tag{1}$$

$$2.\quad \forall \tilde{\boldsymbol{x}} \in \mathcal{D},\ \forall f(\cdot) \in \mathcal{H},\quad f(\tilde{\boldsymbol{x}}) = \langle f(\cdot), K(\cdot, \tilde{\boldsymbol{x}}) \rangle_{\mathcal{H}}, \tag{2}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ stands for the inner product of $\mathcal{H}$.

In the following contents, we use the term 'kernel' instead of 'reproducing kernel' for simplicity.

The Hilbert space $\mathcal{H}$ that has a kernel is called a reproducing kernel Hilbert space (RKHS). Note that kernels are positive definite and symmetric [4]. If a kernel $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ exists, it is unique [4]. Conversely, every positive definite function $K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ has the unique corresponding RKHS [4]. From Eqs. (1) and (2), it immediately follows that $\langle K(\cdot, \boldsymbol{x}), K(\cdot, \tilde{\boldsymbol{x}}) \rangle_{\mathcal{H}_K} = K(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ for any $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathcal{D}$, which plays a crucial role in kernel-based regression problems.

Let $T := \{(\boldsymbol{x}_i, y_i) \mid i \in \{1, \ldots, n\},\ \boldsymbol{x}_i \in \mathcal{D},\ y_i \in \mathbf{R}\}$

Manuscript received July 19, 2021.
Manuscript revised September 15, 2021.
Manuscript publicized October 5, 2021.

[†]The authors are with the Division of Computer Science and Information Technology, Faculty of Information Science and Technology, Hokkaido University, Sapporo-shi, 060–0814 Japan.
a) E-mail: takira@ist.hokudai.ac.jp
b) E-mail: masanari@ist.hokudai.ac.jp
c) E-mail: imai@ist.hokudai.ac.jp

DOI: 10.1587/transinf.2021EDP7156

be a given training data set with n samples, where $\boldsymbol{x}_i$ and y_i denote an input vector and the corresponding output value, satisfying

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i, \tag{3}$$

where $f(\cdot)$ denotes the unknown true function to be estimated and ϵ_i denotes an *i.i.d.* zero-mean additive noise whose variance is (unknown) $\sigma^2 > 0$. The ordinary KRR is formulated as the following problem [1]–[3].

Problem 1: Find a function $\hat{f}(\cdot)$ that minimizes

$$J_1(\hat{f}(\cdot)) := \sum_{i=1}^{n}(y_i - \hat{f}(\boldsymbol{x}_i))^2 + \gamma\|\hat{f}(\cdot)\|^2_{\mathcal{H}_K}, \tag{4}$$

where $\|\cdot\|_{\mathcal{H}_K}$ denotes the induced norm of the RKHS $\mathcal{H}_K$ corresponding to an adopted kernel K, and γ denotes a positive regularization parameter.

The following theorem, called the nonparametric representer theorem, is a well known result on the representation of a solution of a certain class of kernel regression problems.

Theorem 1: [14] Let $c(\cdot) : (\mathcal{D}\times\mathbf{R}^2)^n \to R\cup\{\infty\}$ be an arbitrary cost function, and let $g(\cdot)$ be a strictly monotonically increasing function defined on $[0,\infty)$. Then, any minimizer $\hat{f}(\cdot)\in\mathcal{H}_K$ of the regularized risk functional

$$c((\boldsymbol{x}_1, y_1, \hat{f}(\boldsymbol{x}_1)), \ldots, (\boldsymbol{x}_n, y_n, \hat{f}(\boldsymbol{x}_n))) + g(\|\hat{f}(\cdot)\|_{\mathcal{H}_K}) \tag{5}$$

can be represented by

$$\hat{f}(\cdot) := \sum_{i=1}^{n}\alpha_i K(\cdot, \boldsymbol{x}_i) = (\boldsymbol{g}_X^{(K)}(\cdot))'\boldsymbol{\alpha}, \tag{6}$$

where $\boldsymbol{g}_X^{(K)}(\cdot) := [K(\cdot,\boldsymbol{x}_1), \ldots, K(\cdot,\boldsymbol{x}_n)]' \in \mathbf{R}^n$ and $\boldsymbol{\alpha} := [\alpha_1, \ldots, \alpha_n]' \in \mathbf{R}^n$ with $'$ standing for the transposition operator, and $X := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ being the set of training input vectors.

Since $J_1(\hat{f}(\cdot))$ in Problem 1 agrees to Eq. (5), the solution of Problem 1 can be represented by Eq. (6). It is well known [1]–[3] that the closed-form solution of the minimizer of Eq. (4) with the function model Eq. (6) is given by

$$\hat{f}^{(K,\gamma)}(\cdot) = (\boldsymbol{g}_X^{(K)}(\cdot))'(G_{XX}^{(K)} + \gamma I_n)^{-1}\boldsymbol{y}, \tag{7}$$

where $G_{XX}^{(K)} := (K(\boldsymbol{x}_i,\boldsymbol{x}_j)) \in \mathbf{R}^{n\times n}$ denotes the Gram matrix of the kernel K with X, I_n denotes the identity matrix of degree n, and $\boldsymbol{y} := [y_1, \ldots, y_n]' \in \mathbf{R}^n$.

Note that when $G_{XX}^{(K)}$ is non-singular, Eq. (7) is the unique solution of Problem 1. On the other hand, when $G_{XX}^{(K)}$ is singular, Problem 1 may have many solutions including Eq. (7). Also note that the output vector corresponding to an arbitrary set of test input vectors $Z := \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m\}$, $(\boldsymbol{z}_i \in \mathcal{D})$ is given by

$$\begin{aligned}\hat{\boldsymbol{f}}_Z^{(K,\gamma)} &:= [\hat{f}^{(K,\gamma)}(\boldsymbol{z}_1), \ldots, \hat{f}^{(K,\gamma)}(\boldsymbol{z}_m)]' \\ &= G_{ZX}^{(K)}(G_{XX}^{(K)} + \gamma I_n)^{-1}\boldsymbol{y}\end{aligned} \tag{8}$$

from Eq. (7), where $G_{ZX}^{(K)} := (K(\boldsymbol{z}_i,\boldsymbol{x}_j)) \in \mathbf{R}^{m\times n}$.

3. Kernel Ridge Regression with Autocorrelation Prior and Its relation to Stochastic Linear Estimator

In this section, we review the KRR with the autocorrelation prior introduced in our previous work [5], and discuss its relation to stochastic linear estimators.

We assume that the unknown true function $f(\cdot)$ is a realization of a random process whose autocorrelation function is defined as

$$R(\boldsymbol{x},\tilde{\boldsymbol{x}}) := E_f[f(\boldsymbol{x})f(\tilde{\boldsymbol{x}})], \quad (\boldsymbol{x},\tilde{\boldsymbol{x}}\in\mathcal{D}), \tag{9}$$

where E_f denotes the expectation operator over the random process. Let $Z := \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m\}$ be an arbitrary finite subset of $\mathcal{D}$ and let $\boldsymbol{f}_Z := [f(\boldsymbol{z}_1), \ldots, f(\boldsymbol{z}_m)]'$. Since

$$R_{ZZ} := (R(\boldsymbol{z}_i,\boldsymbol{z}_j)) = (E[f(\boldsymbol{z}_i)f(\boldsymbol{z}_j)]) = E[\boldsymbol{f}_Z\boldsymbol{f}_Z'] \tag{10}$$

is trivially a non-negative definite (*n.n.d.*) matrix, it is concluded that the autocorrelation function $R(\boldsymbol{x},\tilde{\boldsymbol{x}})$, $(\boldsymbol{x},\tilde{\boldsymbol{x}}\in\mathcal{D})$ is also a kernel [15], [16]. Hereafter, we call $R(\boldsymbol{x},\tilde{\boldsymbol{x}})$ the autocorrelation kernel. Since R_{ZZ} is identical to the Gram matrix of the autocorrelation kernel R with an arbitrary finite subset $Z\subset\mathcal{D}$, we use the symbol $G_{ZZ}^{(R)}$ instead of R_{ZZ} in the following contents.

We assume that $f(\boldsymbol{x}_i)$ and ϵ_j are uncorrelated. Let $\boldsymbol{\epsilon} := [\epsilon_1, \ldots, \epsilon_n]'$ and $\boldsymbol{f}_X := [f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)]'$, then the correlation matrix of the training output vector $\boldsymbol{y} = \boldsymbol{f}_X + \boldsymbol{\epsilon}$ is reduced to

$$\begin{aligned}E_{f,\boldsymbol{\epsilon}}[\boldsymbol{y}\boldsymbol{y}'] &= E_{f,\boldsymbol{\epsilon}}[(\boldsymbol{f}_X+\boldsymbol{\epsilon})(\boldsymbol{f}_X+\boldsymbol{\epsilon})'] \\ &= E_f[\boldsymbol{f}_X\boldsymbol{f}_X'] + E_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'] \\ &= G_{XX}^{(R)} + \sigma^2 I_n.\end{aligned} \tag{11}$$

We evaluate the generalization error of the KRR with a model (K,γ) for an arbitrary set of test input vectors $Z\subset\mathcal{D}$ by

$$L(K,\gamma;Z) := \|\hat{\boldsymbol{f}}_Z^{(K,\gamma)} - \boldsymbol{f}_Z\|^2. \tag{12}$$

The following theorem is one of main results in our previous work [5].

Theorem 2: [5] Let K be an arbitrary kernel defined on $\mathcal{D}\times\mathcal{D}$ and let γ be an arbitrary positive constant, then

$$E_{f,\boldsymbol{\epsilon}}L(K,\gamma;Z) \geq E_{f,\boldsymbol{\epsilon}}L(R,\sigma^2;Z) \tag{13}$$

holds for any $Z\subset\mathcal{D}$.

According to Theorem 2, it is concluded that the model (R,σ^2) is optimal for the KRR in terms of the expected generalization error, which agrees to the optimal model (R,σ^2) in the GPR, specified by the conditional expectation based on the training data set.

Next, let us consider the linear estimation model

$$\hat{\boldsymbol{f}}_Z^{(L)} = B\boldsymbol{y}, \tag{14}$$

where $B \in \mathbf{R}^{m\times n}$, and also consider the following problem for this estimation model.

Problem 2: Find the matrix $B \in \mathbf{R}^{m\times n}$ that minimizes

$$J_2(B) := E_{f,\boldsymbol{\epsilon}}\|B\boldsymbol{y} - \boldsymbol{f}_Z\|^2. \tag{15}$$

From Eq. (11) and

$$E_f[\boldsymbol{f}_Z\boldsymbol{f}_Z'] = G_{ZZ}^{(R)},$$
$$E_{f,\boldsymbol{\epsilon}}[\boldsymbol{y}\boldsymbol{f}_Z'] = E_f[\boldsymbol{f}_X\boldsymbol{f}_Z'] = G_{XZ}^{(R)},$$

where $G_{ZZ}^{(R)} := (R(\boldsymbol{z}_i,\boldsymbol{z}_j)) \in \mathbf{R}^{m\times m}$, $G_{XZ}^{(R)} := (R(\boldsymbol{x}_i,\boldsymbol{z}_j)) \in \mathbf{R}^{n\times m}$,

$$\begin{aligned} J_2(B) &:= E_{f,\boldsymbol{\epsilon}}\|B\boldsymbol{y} - \boldsymbol{f}_Z\|^2 \\ &= \mathrm{tr}[B(G_{XX}^{(R)} + \sigma^2 I_n)B' - 2BG_{XZ}^{(R)} + G_{ZZ}^{(R)}] \end{aligned}$$

is obtained and its first order differential is reduced to

$$\begin{aligned} \mathrm{d}J_2(B) &= 2\,\mathrm{tr}[\mathrm{d}B(G_{XX}^{(R)} + \sigma^2 I_n)B' - \mathrm{d}BG_{XZ}^{(R)}] \\ &= 2\,\mathrm{tr}[\mathrm{d}B((G_{XX}^{(R)} + \sigma^2 I_n)B' - G_{XZ}^{(R)})]. \end{aligned}$$

Since $J_2(B)$ is a non-negative quadratic function with respect to B, the stationary point of $J_2(B)$ is its minimizer. Thus, the solution of Problem 2 is immediately obtained by

$$\hat{B} = G_{ZX}^{(R)}(G_{XX}^{(R)} + \sigma^2 I_n)^{-1} \tag{16}$$

and the estimate of $\boldsymbol{f}_Z$ by this optimal matrix $\hat{B}$ is written by

$$\hat{\boldsymbol{f}}_Z^{(L)} = G_{ZX}^{(R)}(G_{XX}^{(R)} + \sigma^2 I_n)^{-1}\boldsymbol{y}, \tag{17}$$

where $G_{ZX}^{(R)} := (G_{XZ}^{(R)})' = (R(\boldsymbol{z}_i,\boldsymbol{x}_j)) \in \mathbf{R}^{m\times n}$, which is identical to Eq. (8) with the optimal model (R,σ^2). Accordingly, it is concluded that the KRR with the optimal model (R,σ^2) is identical to the stochastic linear estimator defined by Problem 2. Note that we do not assume $E_f[\boldsymbol{f}_X] = \boldsymbol{0}_n$ and $E_f[\boldsymbol{f}_Z] = \boldsymbol{0}_m$, where $\boldsymbol{0}_n \in \mathbf{R}^n$ denotes the n-dimensional zero vector, which implies that the Gram matrices are not always reduced to covariance matrices. When $E_f[\boldsymbol{f}_X] = \boldsymbol{0}_n$ and $E_f[\boldsymbol{f}_Z] = \boldsymbol{0}_m$ hold, the above discussion, concerned with a stochastic linear estimator, is the same with that given in [6].

4. Stochastic Affine Estimator and Its Properties

In this section, we discuss a stochastic affine estimator, which is one of the simplest extensions of a stochastic linear estimator. Note that almost all of results of this section is given in [6]. However, we adopt a different way to obtain the same results in order for theoretical analyses concerned with kernel regression problems given in the next section.

Let us consider the following affine estimation model:

$$\hat{\boldsymbol{f}}_Z^{(A)} = B\begin{bmatrix} \boldsymbol{y} \\ 1 \end{bmatrix}, \tag{18}$$

where $B \in \mathbf{R}^{m\times(n+1)}$. As the same with the linear case, we consider the following problem.

Problem 3: Find the matrix $B \in \mathbf{R}^{m\times(n+1)}$ that minimizes

$$J_3(B) := E_{f,\boldsymbol{\epsilon}}\left\|B\begin{bmatrix} \boldsymbol{y} \\ 1 \end{bmatrix} - \boldsymbol{f}_Z\right\|^2. \tag{19}$$

Let $\boldsymbol{\mu}_X := E_f[\boldsymbol{f}_X]$, $\boldsymbol{\mu}_Z := E_f[\boldsymbol{f}_Z]$, and $M := G_{XX}^{(R)} + \sigma^2 I_n$, then

$$E_{f,\boldsymbol{\epsilon}}\left[\begin{bmatrix} \boldsymbol{y} \\ 1 \end{bmatrix}\begin{bmatrix} \boldsymbol{y}' & 1 \end{bmatrix}\right] = \begin{bmatrix} M & \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_X' & 1 \end{bmatrix} =: M_A$$

and

$$E_{f,\boldsymbol{\epsilon}}\left[\begin{bmatrix} \boldsymbol{y} \\ 1 \end{bmatrix}\boldsymbol{f}_Z'\right] = \begin{bmatrix} G_{XZ}^{(R)} \\ \boldsymbol{\mu}_Z' \end{bmatrix} =: Q$$

hold. Therefore

$$\begin{aligned} J_3(B) &:= E_{f,\boldsymbol{\epsilon}}\left\|B\begin{bmatrix} \boldsymbol{y} \\ 1 \end{bmatrix} - \boldsymbol{f}_Z\right\|^2 \\ &= \mathrm{tr}\left[BM_AB' - 2BQ + G_{ZZ}^{(R)}\right] \end{aligned}$$

is obtained and its first order differential is reduced to

$$\begin{aligned} \mathrm{d}J_3(B) &= 2\,\mathrm{tr}\left[\mathrm{d}BM_AB' - \mathrm{d}BQ\right] \\ &= 2\,\mathrm{tr}\left[\mathrm{d}B\left(M_AB' - Q\right)\right]. \end{aligned}$$

Theorem 3: The matrix M_A is non-singular.

Proof The covariance matrix Σ_X of $\boldsymbol{f}_X$ is represented by

$$\begin{aligned} \Sigma_X &:= E_f[(\boldsymbol{f}_X - \boldsymbol{\mu}_X)(\boldsymbol{f}_X - \boldsymbol{\mu}_X)'] \\ &= E_f[\boldsymbol{f}_X\boldsymbol{f}_X'] - \boldsymbol{\mu}_X\boldsymbol{\mu}_X' \\ &= G_{XX}^{(R)} - \boldsymbol{\mu}_X\boldsymbol{\mu}_X'. \end{aligned} \tag{20}$$

Therefore,

$$M = \Sigma_X + \sigma^2 I_n + \boldsymbol{\mu}_X\boldsymbol{\mu}_X'$$

and

$$M_A = T_1 + T_2$$

are obtained with

$$T_1 := \begin{bmatrix} \Sigma_X + \sigma^2 I_n & \boldsymbol{0}_n \\ \boldsymbol{0}_n' & 0 \end{bmatrix}, \quad T_2 := \begin{bmatrix} \boldsymbol{\mu}_X \\ 1 \end{bmatrix}\begin{bmatrix} \boldsymbol{\mu}_X' & 1 \end{bmatrix}.$$

Let $[\boldsymbol{u}' \;\; v]' \in \mathbf{R}^{n+1}$ be an arbitrary vector in the null space of M_A, written as $\mathcal{N}(M_A)$, then

$$\begin{bmatrix} \boldsymbol{u}' & v \end{bmatrix}T_1\begin{bmatrix} \boldsymbol{u} \\ v \end{bmatrix} = 0, \tag{21}$$

$$\begin{bmatrix} \boldsymbol{u}' & v \end{bmatrix}T_2\begin{bmatrix} \boldsymbol{u} \\ v \end{bmatrix} = 0 \tag{22}$$

must hold since T_1 and T_2 are *n.n.d.* Therefore, Eq. (21) immediately yields $\boldsymbol{u} = \boldsymbol{0}_n$ since $\Sigma_X + \sigma^2 I_n$ is non-singular, and then Eq. (22) yields $v = 0$, which implies $\mathcal{N}(M_A) = \{\boldsymbol{0}_{n+1}\}$. □

Since $J_3(B)$ is a non-negative quadratic function with respect to B, the stationary point of $J_3(B)$ is the minimizer of $J_3(B)$. Thus, the solution of Problem 3 is immediately obtained by

$$\hat{B} = Q' M_A^{-1}. \tag{23}$$

By substituting Eq. (23) into Eq. (18), the estimate of $\boldsymbol{f}_Z$ by the solution of Problem 3 is reduced to

$$\begin{aligned}\hat{\boldsymbol{f}}_Z^{(A)} &= Q' M_A^{-1} \begin{bmatrix} \boldsymbol{y} \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} G_{ZX}^{(R)} & \boldsymbol{\mu}_Z \end{bmatrix} \\ &\quad \times \begin{bmatrix} G_{XX}^{(R)} + \sigma^2 I_n & \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_X' & 1 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{y} \\ 1 \end{bmatrix}. \end{aligned} \tag{24}$$

5. Kernel Regressors Equivalent to Stochastic Affine Estimators

In this section, we discuss the kernel regression problem, whose solution is reduced to the stochastic affine estimator $\hat{\boldsymbol{f}}_Z^{(A)}$ in Eq. (24).

The function with the model (K, γ) that gives Eq. (24) by substituting the points in Z can be represented by

$$\begin{aligned}\hat{f}^{(K,\gamma)}(\cdot) &= \begin{bmatrix} (\boldsymbol{g}_X^{(K)}(\cdot))' & \mu(\cdot) \end{bmatrix} \\ &\quad \times \begin{bmatrix} G_{XX}^{(K)} + \gamma I_n & \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_X' & 1 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{y} \\ 1 \end{bmatrix}, \end{aligned} \tag{25}$$

where $\mu(\cdot)$ is a certain function, that yields $\boldsymbol{\mu}_X := [\mu(\boldsymbol{x}_1) \ldots, \mu(\boldsymbol{x}_n)]'$, corresponding to the mean vector in the stochastic affine estimator. Thus, we consider the following function model:

$$\hat{f}(\cdot) := (\boldsymbol{g}_X^{(K)}(\cdot))' \boldsymbol{\alpha} + \beta \mu(\cdot), \tag{26}$$

where $\boldsymbol{\alpha} \in \mathbf{R}^n$ and $\beta \in \mathbf{R}$.

Let $K_\mu(\boldsymbol{x}, \tilde{\boldsymbol{x}}) := \mu(\boldsymbol{x})\mu(\tilde{\boldsymbol{x}})$, $(\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathcal{D})$, then $K_\mu(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ is the kernel whose corresponding RKHS is consisting of $a\mu(\cdot)$, $(a \in \mathbf{R})$ as shown in [13], which trivially implies $\mu(\cdot) \in \mathcal{H}_{K_\mu}$. Note that since

$$\begin{aligned}\mu(\boldsymbol{x}) &= \langle \mu(\cdot), K_\mu(\cdot, \boldsymbol{x}) \rangle_{\mathcal{H}_{K_\mu}} \\ &= \langle \mu(\cdot), \mu(\cdot)\mu(\boldsymbol{x}) \rangle_{\mathcal{H}_{K_\mu}} \\ &= \mu(\boldsymbol{x}) \langle \mu(\cdot), \mu(\cdot) \rangle_{\mathcal{H}_{K_\mu}} \\ &= \mu(\boldsymbol{x}) \|\mu(\cdot)\|_{\mathcal{H}_{K_\mu}}^2 \end{aligned}$$

holds for any $\boldsymbol{x} \in \mathcal{D}$, we have $\|\mu(\cdot)\|_{\mathcal{H}_{K_\mu}}^2 = 1$.

Theorem 4: [4] If K_i is the reproducing kernel of the class F_i with the norm $\|\cdot\|_i$, then $K = K_1 + K_2$ is the reproducing kernel of the class F of all functions $f(\cdot) = f_1(\cdot) + f_2(\cdot)$ with $f_i(\cdot) \in F_i$, and with the norm defined by

$$\|f(\cdot)\|^2 = \min \left[\|f_1(\cdot)\|_1^2 + \|f_2(\cdot)\|_2^2 \right], \tag{27}$$

the minimum taken for all the decompositions $f(\cdot) = f_1(\cdot) + f_2(\cdot)$ with $f_i(\cdot) \in F_i$.

Hereafter, we assume that the kernel, used in the following contents, is of the form:

$$K(\boldsymbol{x}, \tilde{\boldsymbol{x}}) := K_c(\boldsymbol{x}, \tilde{\boldsymbol{x}}) + K_\mu(\boldsymbol{x}, \tilde{\boldsymbol{x}}), \tag{28}$$

where K_c is a certain kernel, whose Gram matrix with X corresponds to a covariance matrix Σ_X in the stochastic affine estimator as in Eq. (20). We call K_c the covariance part of K and also call K_μ the mean part of K. It should be noted that $\mu(\cdot) \in \mathcal{H}_K$ holds due to Theorem 4 and the fact that $\mu(\cdot) \in \mathcal{H}_{K_\mu}$. We assume $\mu(\cdot) \notin \mathcal{H}_{K_c}$[†], which implies $\mathcal{H}_{K_c} \cap \mathcal{H}_{K_\mu} = \{0\}$, and then $\|\mu(\cdot)\|_{\mathcal{H}_K}^2 = 1$ is obtained, since the decomposition in Theorem 4 is unique. We also assume that $\mu(\cdot) \notin \mathcal{L}_X^{(K)} := \mathrm{span}\{K(\cdot, \boldsymbol{x}_i) \mid i \in \{1, \ldots, n\}\}$, since when $\mu(\cdot) \in \mathcal{L}_X^{(K)}$ holds, Eq. (26) is reduced to Eq. (6). Note that the Gram matrix of K with X can be represented as

$$G_{XX}^{(K)} = G_{XX}^{(K_c)} + G_{XX}^{(K_\mu)} = G_{XX}^{(K_c)} + \boldsymbol{\mu}_X \boldsymbol{\mu}_X'.$$

Here, we define some subsets of $\mathcal{H}_K$. Let $\mathcal{L}_\mu := \mathrm{span}\{\mu(\cdot)\}$ and let $\mathcal{L}_\mu^\perp$ be its orthogonal complement, then, we define the linear manifold

$$\mathcal{M}_\mu := \{\mu^\perp(\cdot) + \mu(\cdot) \mid \mu^\perp(\cdot) \in \mathcal{L}_\mu^\perp\} \subset \mathcal{H}_K.$$

Under these preparations, we define the following problem.

Problem 4: Find the function $\hat{f}(\cdot) \in \mathcal{M}_\mu$ that minimizes

$$J_4(\hat{f}(\cdot)) := \sum_{i=1}^n (y_i - \hat{f}(\boldsymbol{x}_i))^2 + \gamma \|\hat{f}(\cdot)\|_{\mathcal{H}_K}^2 \tag{29}$$

with a positive constant γ.

It should be noted that the constraint

$$\hat{f}(\cdot) \in \mathcal{M}_\mu \tag{30}$$

can be represented as

$$\mu(\cdot) = P_\mu \hat{f}(\cdot), \tag{31}$$

where P_μ denotes the orthogonal projector onto $\mathcal{L}_\mu$, since $\hat{f}(\cdot) \in \mathcal{M}_\mu$ can be represented as

$$\hat{f}(\cdot) = f^\perp(\cdot) + \mu(\cdot)$$

with $f^\perp(\cdot) \in \mathcal{L}_\mu^\perp$ and

$$P_\mu \hat{f}(\cdot) = P_\mu f^\perp(\cdot) + P_\mu \mu(\cdot) = 0 + P_\mu \mu(\cdot) = \mu(\cdot)$$

is immediately followed.

The first issue to be resolved is whether the solution of Problem 4 can be represented by Eq. (26) or not. In [14], the representer theorem concerned with the function model Eq. (26), which is called the semiparametric representer theorem, is given as follows.

Theorem 5: [14] In addition to the assumptions of Theorem 1, we assume that a set of n_P real-valued function

[†]This is the only arbitrary assumption that the framework of the stochastic affine estimators has no counterpart.

$\{\psi_p(\cdot)\}$, $(p \in \{1,\ldots,n_P\})$ defined on $\mathcal{D}$ with the property that the matrix $(\psi_p(\boldsymbol{x}_i)) \in \mathbf{R}^{n_P \times n}$ has rank n_P is given. Then, any $\hat{f}(\cdot) := f(\cdot) + h(\cdot)$ with $f(\cdot) \in \mathcal{H}_K$ and $h(\cdot) \in \mathrm{span}\{\psi_1(\cdot),\ldots,\psi_{n_P}(\cdot)\}$, minimizing the regularized risk functional

$$c((\boldsymbol{x}_1, y_1, \hat{f}(\boldsymbol{x}_1)),\ldots,(\boldsymbol{x}_n, y_n, \hat{f}(\boldsymbol{x}_n))) + g(\|f(\cdot)\|_{\mathcal{H}_K}) \quad (32)$$

can be represented by

$$\hat{f}(\cdot) := \sum_{i=1}^{n} \alpha_i K(\cdot, \boldsymbol{x}_i) + \sum_{p=1}^{n_P} \beta_p \psi_p(\cdot) \quad (33)$$

with unique coefficients $\beta_p \in \mathbf{R}$ for all $p \in \{1,\ldots,n_P\}$.

The function model Eq. (26) agrees to Eq. (33) with $n_P = 1$. However, the optimization criterion and the constraint in Problem 4 does not agree to Eq. (32), which implies that Theorem 5 can not be applied to problem 4. In order to positively resolve this issue, we give the following theorem.

Theorem 6: Any function $\hat{f}(\cdot) \in \mathcal{M}_\mu$ that minimizes Eq. (29) with a positive γ can be represented by Eq. (26).

Proof Let $v(\cdot) \in \mathcal{H}_K$ be an arbitrary function orthogonal to $\mathcal{L}_\mu + \mathcal{L}_X^{(K)}$. Without loss of generality, any function $\hat{f}(\cdot) \in \mathcal{H}_K$ can be represented by

$$\hat{f}(\cdot) = (\boldsymbol{g}_X^{(K)}(\cdot))'\boldsymbol{\alpha} + \beta\mu(\cdot) + v(\cdot).$$

Since $v(\cdot) \in \left(\mathcal{L}_\mu + \mathcal{L}_X^{(K)}\right)^\perp$, we have $\langle \mu(\cdot), v(\cdot)\rangle_{\mathcal{H}_K} = 0$ and $\langle v(\cdot), K(\cdot, \boldsymbol{x}_i)\rangle_{\mathcal{H}_K} = 0$ for any $i \in \{1,\ldots,n\}$. The former leads

$$P_\mu \hat{f}(\cdot) = P_\mu((\boldsymbol{g}_X^{(K)}(\cdot))'\boldsymbol{\alpha} + \beta\mu(\cdot)),$$

which implies that the constraint Eq. (30) is independent of $v(\cdot)$. Similarly, the latter leads

$$\begin{aligned}\hat{f}(\boldsymbol{x}_i) &= \langle \hat{f}(\cdot), K(\cdot, \boldsymbol{x}_i)\rangle_{\mathcal{H}_K} \\ &= \langle (\boldsymbol{g}_X^{(K)}(\cdot))'\boldsymbol{\alpha} + \beta\mu(\cdot) + v(\cdot), K(\cdot, \boldsymbol{x}_i)\rangle_{\mathcal{H}_K} \\ &= \langle (\boldsymbol{g}_X^{(K)}(\cdot))'\boldsymbol{\alpha} + \beta\mu(\cdot), K(\cdot, \boldsymbol{x}_i)\rangle_{\mathcal{H}_K},\end{aligned}$$

which implies that the first term of Eq. (29) is also independent of $v(\cdot)$. The second term of Eq. (29) satisfies

$$\begin{aligned}&\gamma\|\hat{f}(\cdot)\|_{\mathcal{H}_K}^2 = \gamma\langle \hat{f}(\cdot), \hat{f}(\cdot)\rangle_{\mathcal{H}_K} \\ &= \gamma\langle (\boldsymbol{g}_X^{(K)}(\cdot))'\boldsymbol{\alpha} + \beta\mu(\cdot) + v(\cdot), \\ &\qquad (\boldsymbol{g}_X^{(K)}(\cdot))'\boldsymbol{\alpha} + \beta\mu(\cdot) + v(\cdot)\rangle_{\mathcal{H}_K} \\ &= \gamma\|(\boldsymbol{g}_X^{(K)}(\cdot))'\boldsymbol{\alpha} + \beta\mu(\cdot)\|_{\mathcal{H}_K}^2 + \gamma\|v(\cdot)\|_{\mathcal{H}_K}^2 \\ &\geq \gamma\|(\boldsymbol{g}_X^{(K)}(\cdot))'\boldsymbol{\alpha} + \beta\mu(\cdot)\|_{\mathcal{H}_K}^2\end{aligned}$$

since $\gamma > 0$, and equality is attained if and only if $v(\cdot) = 0$. Accordingly, in order for $\hat{f}(\cdot)$ to be a minimizer of Eq. (29), $v(\cdot) = 0$ must hold, which concludes the proof. □

According to Theorem 6, it is enough to consider the function model Eq. (26) as a candidate of the solution of Problem 4.

Since $\hat{f}(\cdot)$ in Problem 4 is specified by $\boldsymbol{\alpha}$ and β only, we use the notation $J_4(\boldsymbol{\alpha}, \beta)$ as the criterion of Problem 4 instead of $J_4(\hat{f}(\cdot))$. Let $S_i(\boldsymbol{\alpha}, \beta)$, $(i \in \{1, 2\})$ be the i-th term of Eq. (29). It is easy to show that

$$S_1(\boldsymbol{\alpha}, \beta) = \left\| \boldsymbol{y} - \begin{bmatrix} G_{XX}^{(K)} & \boldsymbol{\mu}_X \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \beta \end{bmatrix} \right\|^2.$$

Also, we have

$$\begin{aligned}S_2(\boldsymbol{\alpha}, \beta) &= \gamma\|\hat{f}(\cdot)\|_{\mathcal{H}_K}^2 \\ &= \gamma\|(\boldsymbol{g}_X^{(K)}(\cdot))'\boldsymbol{\alpha} + \beta\mu(\cdot)\|_{\mathcal{H}_K}^2 \\ &= \gamma(\boldsymbol{\alpha}' G_{XX}^{(K)}\boldsymbol{\alpha} + 2\beta\boldsymbol{\alpha}'\boldsymbol{\mu}_X + \beta^2) \\ &= \gamma \begin{bmatrix} \boldsymbol{\alpha}' & \beta \end{bmatrix} \begin{bmatrix} G_{XX}^{(K)} & \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_X' & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \beta \end{bmatrix}\end{aligned}$$

for $\hat{f}(\cdot)$ specified by the function model Eq. (26).

As mentioned before, the linear manifold constraint Eq. (30) is identical to Eq. (31), which can be also represented as

$$\begin{aligned}\mu(\cdot) &= P_\mu \hat{f}(\cdot) \\ &= \langle \hat{f}(\cdot), \mu(\cdot)\rangle_{\mathcal{H}_K}\mu(\cdot) \\ &= \langle (\boldsymbol{g}_X^{(K)}(\cdot))'\boldsymbol{\alpha} + \beta\mu(\cdot), \mu(\cdot)\rangle_{\mathcal{H}_K}\mu(\cdot) \\ &= (\boldsymbol{\mu}_X'\boldsymbol{\alpha} + \beta)\,\mu(\cdot). \end{aligned} \quad (34)$$

Since $\|\mu(\cdot)\|_{\mathcal{H}_K}^2 = 1 \neq 0$ holds, Eq. (34) is identical to

$$1 = \boldsymbol{\mu}_X'\boldsymbol{\alpha} + \beta, \quad (35)$$

which can be also represented as

$$1 = \begin{bmatrix} \boldsymbol{\mu}_X' & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \beta \end{bmatrix} \quad (36)$$

and

$$\begin{bmatrix} \boldsymbol{\mu}_X \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_X \\ 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}_X' & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \beta \end{bmatrix}, \quad (37)$$

since $[\boldsymbol{\mu}_X'\ 1]'$ is an $(n+1) \times 1$ full column rank matrix.

The following theorem is the main result of this paper.

Theorem 7: Eq. (25) is a solution of Problem 4.

Proof Let

$$\mathscr{L}(\boldsymbol{\alpha}, \beta, \lambda) = \sum_{i=1}^{2} S_i(\boldsymbol{\alpha}, \beta) - 2\lambda(\boldsymbol{\mu}_X'\boldsymbol{\alpha} + \beta - 1) \quad (38)$$

be the Lagrangian function induced from Problem 4 with a Lagrange multiplier λ. Since

$$\begin{aligned}&\mathscr{L}(\boldsymbol{\alpha}, \beta, \lambda) \\ &= \|\boldsymbol{y}\|^2 + \begin{bmatrix} \boldsymbol{\alpha}' & \beta \end{bmatrix} \begin{bmatrix} G_{XX}^{(K)} \\ \boldsymbol{\mu}_X' \end{bmatrix} \begin{bmatrix} G_{XX}^{(K)} & \boldsymbol{\mu}_X \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \beta \end{bmatrix} \\ &\quad -2 \begin{bmatrix} \boldsymbol{\alpha}' & \beta \end{bmatrix} \begin{bmatrix} G_{XX}^{(K)} \\ \boldsymbol{\mu}_X' \end{bmatrix} \boldsymbol{y}\end{aligned}$$

$$
\begin{aligned}
&+\gamma\begin{bmatrix}\boldsymbol{\alpha}' & \beta\end{bmatrix}\begin{bmatrix}G_{XX}^{(K)} & \boldsymbol{\mu}_X\\ \boldsymbol{\mu}_X' & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ \beta\end{bmatrix}\\
&-2\lambda\left(\begin{bmatrix}\boldsymbol{\alpha}' & \beta\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}_X\\ 1\end{bmatrix}-1\right)
\end{aligned}
$$

holds, its first order differential with respect to $[\boldsymbol{\alpha}'\ \beta]'$ is reduced to

$$
\begin{aligned}
&\mathrm{d}\mathscr{L}(\boldsymbol{\alpha},\beta,\lambda)\\
&=2\mathrm{d}\begin{bmatrix}\boldsymbol{\alpha}' & \beta\end{bmatrix}\begin{bmatrix}G_{XX}^{(K)}\\ \boldsymbol{\mu}_X'\end{bmatrix}\begin{bmatrix}G_{XX}^{(K)} & \boldsymbol{\mu}_X\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ \beta\end{bmatrix}\\
&\quad-2\mathrm{d}\begin{bmatrix}\boldsymbol{\alpha}' & \beta\end{bmatrix}\begin{bmatrix}G_{XX}^{(K)}\\ \boldsymbol{\mu}_X'\end{bmatrix}\boldsymbol{y}\\
&\quad+2\gamma\mathrm{d}\begin{bmatrix}\boldsymbol{\alpha}' & \beta\end{bmatrix}\begin{bmatrix}G_{XX}^{(K)} & \boldsymbol{\mu}_X\\ \boldsymbol{\mu}_X' & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ \beta\end{bmatrix}\\
&\quad-2\lambda\mathrm{d}\begin{bmatrix}\boldsymbol{\alpha}' & \beta\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}_X\\ 1\end{bmatrix}\\
&=2\mathrm{d}\begin{bmatrix}\boldsymbol{\alpha}' & \beta\end{bmatrix}\begin{bmatrix}G_{XX}^{(K)}\\ \boldsymbol{\mu}_X'\end{bmatrix}\begin{bmatrix}G_{XX}^{(K)} & \boldsymbol{\mu}_X\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ \beta\end{bmatrix}\\
&\quad+2\gamma\mathrm{d}\begin{bmatrix}\boldsymbol{\alpha}' & \beta\end{bmatrix}\begin{bmatrix}G_{XX}^{(K)} & \boldsymbol{\mu}_X\\ \boldsymbol{\mu}_X' & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ \beta\end{bmatrix}\\
&\quad-2\mathrm{d}\begin{bmatrix}\boldsymbol{\alpha}' & \beta\end{bmatrix}\begin{bmatrix}G_{XX}^{(K)} & \boldsymbol{\mu}_X\\ \boldsymbol{\mu}_X' & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{y}\\ \lambda\end{bmatrix}. \qquad (39)
\end{aligned}
$$

Since $J_4(\boldsymbol{\alpha},\beta)$ is a non-negative quadratic function with respect to $[\boldsymbol{\alpha}'\beta]'$, Eq. (38) is also a lower-bounded quadratic function with respect to $[\boldsymbol{\alpha}'\ \beta]'$ under the constraint Eq. (30). Thus, stationary points of Eq. (38) is its minimizer. Therefore,

$$
\begin{aligned}
&\left(\begin{bmatrix}G_{XX}^{(K)}\\ \boldsymbol{\mu}_X'\end{bmatrix}\begin{bmatrix}G_{XX}^{(K)} & \boldsymbol{\mu}_X\end{bmatrix}+\gamma\begin{bmatrix}G_{XX}^{(K)} & \boldsymbol{\mu}_X\\ \boldsymbol{\mu}_X' & 1\end{bmatrix}\right)\begin{bmatrix}\boldsymbol{\alpha}\\ \beta\end{bmatrix}\\
&-\begin{bmatrix}G_{XX}^{(K)} & \boldsymbol{\mu}_X\\ \boldsymbol{\mu}_X' & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{y}\\ \lambda\end{bmatrix}=\mathbf{0}_{n+1} \qquad (40)
\end{aligned}
$$

must hold, under the constraint Eq. (30). Note that Eq. (30) can be represented by

$$
\begin{bmatrix}\boldsymbol{\mu}_X\\ 1\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}_X' & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ \beta\end{bmatrix}-\begin{bmatrix}\boldsymbol{\mu}_X\\ 1\end{bmatrix}=\mathbf{0}_{n+1} \qquad (41)
$$

from Eq. (37). Since Eqs. (40) and (41) must hold simultaneously, the sum of their left sides must be zero, which is represented by

$$
\begin{aligned}
&\left(\begin{bmatrix}G_{XX}^{(K)} & \boldsymbol{\mu}_X\\ \boldsymbol{\mu}_X' & 1\end{bmatrix}^2+\gamma\begin{bmatrix}G_{XX}^{(K)} & \boldsymbol{\mu}_X\\ \boldsymbol{\mu}_X' & 1\end{bmatrix}\right)\begin{bmatrix}\boldsymbol{\alpha}\\ \beta\end{bmatrix}\\
&-\begin{bmatrix}G_{XX}^{(K)} & \boldsymbol{\mu}_X\\ \boldsymbol{\mu}_X' & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{y}\\ \lambda+1\end{bmatrix}=\mathbf{0}_{n+1}, \qquad (42)
\end{aligned}
$$

that can be also represented as

$$
\begin{aligned}
&\begin{bmatrix}G_{XX}^{(K)} & \boldsymbol{\mu}_X\\ \boldsymbol{\mu}_X' & 1\end{bmatrix}\begin{bmatrix}G_{XX}^{(K)}+\gamma I_n & \boldsymbol{\mu}_X\\ \boldsymbol{\mu}_X' & 1+\gamma\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ \beta\end{bmatrix}\\
&-\begin{bmatrix}G_{XX}^{(K)} & \boldsymbol{\mu}_X\\ \boldsymbol{\mu}_X' & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{y}\\ \lambda+1\end{bmatrix}=\mathbf{0}_{n+1}. \qquad (43)
\end{aligned}
$$

Therefore, the solution of the linear equation

$$
\begin{bmatrix}G_{XX}^{(K)}+\gamma I_n & \boldsymbol{\mu}_X\\ \boldsymbol{\mu}_X' & 1+\gamma\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ \beta\end{bmatrix}-\begin{bmatrix}\boldsymbol{y}\\ \lambda+1\end{bmatrix}=\mathbf{0}_{n+1} \qquad (44)
$$

is a stationary point of Eq. (38). Since the last row of Eq. (44) is reduced to

$$
\begin{aligned}
&\boldsymbol{\mu}_X'\boldsymbol{\alpha}+(1+\gamma)\beta-(\lambda+1)\\
&=1+\gamma\beta-(\lambda+1)=\gamma\beta-\lambda=0
\end{aligned}
$$

due to the constraint Eq. (35), $\lambda=\gamma\beta$ is immediately followed. Accordingly, the second term of the left side of Eq. (44) can be represented by

$$
\begin{bmatrix}\boldsymbol{y}\\ \gamma\beta+1\end{bmatrix}=\begin{bmatrix}\boldsymbol{y}\\ 1\end{bmatrix}+\begin{bmatrix}O_n & \mathbf{0}_n\\ \mathbf{0}_n' & \gamma\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ \beta\end{bmatrix}, \qquad (45)
$$

where O_n stands for the $n\times n$ zero matrix, and substituting it to Eq. (44) yields

$$
\begin{bmatrix}G_{XX}^{(K)}+\gamma I_n & \boldsymbol{\mu}_X\\ \boldsymbol{\mu}_X' & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}\\ \beta\end{bmatrix}-\begin{bmatrix}\boldsymbol{y}\\ 1\end{bmatrix}=\mathbf{0}_{n+1}. \qquad (46)
$$

It is easy to show that the matrix

$$
M_B:=\begin{bmatrix}G_{XX}^{(K)}+\gamma I_n & \boldsymbol{\mu}_X\\ \boldsymbol{\mu}_X' & 1\end{bmatrix}
$$

is non-singular by Theorem 3. Thus,

$$
\begin{bmatrix}\boldsymbol{\alpha}\\ \beta\end{bmatrix}=\begin{bmatrix}G_{XX}^{(K)}+\gamma I_n & \boldsymbol{\mu}_X\\ \boldsymbol{\mu}_X' & 1\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{y}\\ 1\end{bmatrix} \qquad (47)
$$

surely exists and it is a minimizer of Eq. (38) under the constraint Eq. (30) and then Eq. (25) is a solution of Problem 4. □

According to Theorem 7, it is concluded that the model (R,σ^2) is also optimal for Problem 4 with a given function $\mu(\cdot)$ in terms of the expected squared error under the autocorrelation prior, as the same with the ordinary KRR. Note that when the matrix

$$
M_C:=\begin{bmatrix}G_{XX}^{(K)} & \boldsymbol{\mu}_X\\ \boldsymbol{\mu}_X' & 1\end{bmatrix}
$$

is non-singular, Eq. (47) gives the unique solution of Problem 4. On the other hand, when M_C is singular, problem 4 may have many solutions including the learning result based on Eq. (47) as the same with the ordinary KRR.

Here, we give some remarks on the optimization criterion of Problem 4. As the same with the ordinary KRR, $S_1(\boldsymbol{\alpha},\beta)$ in $J_4(\boldsymbol{\alpha},\beta)$ represents the fidelity of the estimated function to the given training data set. However, the minimizer of $S_1(\boldsymbol{\alpha},\beta)$ has remaining degree of freedom caused by the singularity of the coefficient matrix in the normal equation obtained from the first order differential of $S_1(\boldsymbol{\alpha},\beta)$. The solution Eq. (47) is obtained by vanishing the degree of freedom by the linear manifold constraint Eq. (30), instead of the minimum norm constraint, which is usually adopted

in solving an underdetermined linear system. It seems that $S_2(\alpha,\beta)$ in $J_4(\alpha,\beta)$ plays the same role of the regularization term in Problem 1. However, it should be noted that the effect of the regularization parameter is restricted to the Gram matrix concerned with the term $(\boldsymbol{g}_X^{(K)}(\cdot))'\alpha$ in the function model Eq. (26) due to the linear manifold constraint. Therefore, it is concluded that the linear manifold constraint $\hat{f}(\cdot) \in \mathcal{M}_\mu$ is crucial in obtaining the kernel-based regressor Eq. (25)

It is expected that replacing $S_2(\alpha,\beta)$ in $J_4(\alpha,\beta)$ to another alternatives yields various kernel regressors represented as affine estimators together with the linear manifold constraint, and our formulation may be useful to construct a kernel-based regression problem with multiple given functions instead of a single $\mu(\cdot)$, while these extensions can not be obtained in the framework of the GPR due to its principle.

6. Conclusion

In this paper, we discussed the kernel-based regression problems by considering the relation to stochastic estimators, and obtained the formulation, which is equivalent to stochastic affine estimators, and gave the interpretation of it. Moreover, we gave a novel representer theorem concerned with the formulation. It is expected that our result give a novel aspect for further extensions of kernel-based regression problems.

Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Number JP20H04238.

References

[1] J. Shawe-Taylor and N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, 2004.
[2] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, Cambridge, 2000.
[3] K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," IEEE Trans. Neural Netw., vol.12, no.2, pp.181–201, 2001.
[4] N. Aronszajn, "Theory of reproducing kernels," Trans. American Mathematical Society, vol.68, no.3, pp.337–404, 1950.
[5] A. Tanaka and H. Imai, "Kernel ridge regression with autocorrelation prior: Optimal model and cross-validation," 2020 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2020), pp.3867–3871, 2020.
[6] T. Kailath, A.H. Sayed, and B. Hassibi, Linear Estimation, Prentice-Hall, Upper Saddle River, NJ, 2000.
[7] C.E. Rasmussen, "Gaussian processes in machine learning," Summer School on Machine Learning, Lecture Notes in Computer Science, vol.3176, pp.63–71, Springer 2004.
[8] C.E. Rasmussen and C.K.I. Williams, Gaussian processes for machine learning, MIT Press, Cambridge, MA, 2006.
[9] K.P. Murphy, Machine learning: A probabilistic perspective, MIT Press, 2012.
[10] M. Kanagawa, P. Hennig, D. Sejdinovic, and B.K. Sriperumbudur, "Gaussian process and kernel methods: A review on connections and equivalences," arXiv:1807.02582 [stat.ML], 2018.
[11] A. Tanaka, H. Imai, and M. Miyakoshi, "Kernel-induced sampling theorem," IEEE Trans. Signal Process., vol.58, no.7, pp.3569–3577, 2010.
[12] J. Mercer, "Functions of positive and negative type, and their connection with the theory of integral equations," Philosophical Trans. Royal Society A, vol.209, no.441-458, pp.415–446, 1909.
[13] S. Saitoh, Integral Transforms, Reproducing Kernels and Their Applications, Addison Wesley Longman Ltd, UK, 1997.
[14] B. Schölkopf, R. Herbrich, and A. Smola, "A generalized representer theorem," Computational Learning Theory, Lecture Notes in Computer Science, vol.2111, pp.416–426, Springer, 2001.
[15] R. Kong and B. Zhang, "Autocorrelation kernel functions for support vector machines," The 3rd International Conference on Natural Computation, 2007.
[16] Y. Saito and A. Tanaka, "Optimal kernel in kernel regression problems with autocorrelation prior," 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp.969–972, IEEE, 2017.

Akira Tanaka received the D.E. from Hokkaido University in 2000. He joined the Faculty of Information Science and Technology, Hokkaido University. His research interests include image processing, acoustic signal processing, and machine learning theory.

Masanari Nakamura received the D.S. from Hokkaido University in 2018. He joined the Faculty of Information Science and Technology, Hokkaido University. His research interests include acoustic signal processing, positioning, and target tracking.

Hideyuki Imai received the D.E. from Hokkaido University in 1999. He joined the Graduate School of Information Science and Technology, Hokkaido University. His research interests include statistical inference.