2023 European Control Conference (ECC)
June 13-16, 2023. Bucharest, Romania

Distributed Regression by Two Agents from Noisy Data*

Aneesh Raghavan and Karl Henrik Johansson

***Abstract*—We consider the problem of learning functions by two agents and a fusion center from noisy data. True data comprises of samples of an independent variable (input) and the corresponding value of a dependent variable (output) collectively labeled as (input, output) data. The data received by the agents, both the input and output data, are corrupted by noise. The objective of the agents is to learn a mapping from the true input to the true output. We formulate a general regression problem for the agents followed by the least squares regression (LS) problem. We prove a stochastic representer theorem for the general regression problem and subsequently solve the LS problem. The functions learned by the agents are transmitted to the fusion center where an optimization problem is formulated to fuse the functions together, which is then declared as the mapping. As an example, the methodology developed has been applied to the data generated from a transcendental function.**

I. Introduction

Data driven methods for control, communication and signal processing have gained importance in the recent past due availability of data and computational power. Data is often available from multiple sources; e.g. , different kinds of sensors, same sensor different capabilities, etc. Communicating large quantities of data could be expensive. Furthermore, processing lot of data simultaneously could be computationally unfeasible. In some applications like IoT, communication of data is not allowed due to privacy issues. Suppose we consider a multi-agent learning problem where input-output data are collected by multiple agents with the objective of learning a mapping from input to output. Conditions listed before (large quantities of data, privacy, etc.) call for processing data locally at each agent and communicating some key features or functions learned from the data to a fusion center. At the fusion center, the functions are fused taking into consideration the given learning problem. Such learning schemes are referred to as distributed learning schemes.

Federated learning (FL) is a distributed learning approach that has received significant attention in the last couple of years, [1], [2]. FL has been applied to problems in variety fields including IoT, smart cities, and heath care. Some of the benefits of FL are enhanced data privacy, low network latency and enhanced learning quality, [3]. There are different kinds of FL, horizontal FL, vertical FL and hybrid FL. In horizontal FL, the agents have minimal intersection of the sample space but the same feature space; in vertical the agents have different feature space but the same sample space; In hybrid FL both the sample space and feature space are different. In majority of the horizontal FL literature, averages or weighted averages of the functions received from different agents is considered at the fusion center. Recently FL methods have been introduced to robotics and autonomy as well, [4]. Kernel based learning methods is a well studied topic in A.I and signal processing, [5], [6]. Kernel based regression and classification have been applied to various problems, [7], [8], [9] including Gaussian process classification, landslide identification, etc. Kernel based learning methods have also been applied to signal processing applications, [10], [11], [12]. Distributed regression has been studied from different perspectives in [13] and [14].

Our contribution to this paper is the following problem. There are two agents, agent 1 and agent 2. Each agent receives samples of an independent variable (input) corrupted by noise and the corresponding value of the dependent variable (output at true input value) corrupted by noise. From an operational standpoint, the dependency of the dependent variable on the independent variable is the same for both the agents. The objective of each agent is to find a function that best captures dependency from true input data to true output data captured by it, that is the data without noise in it. To this end, we formulate a general regression problem with noisy data in a Reproducing Kernel Hilbert Space (RKHS). We prove that the solution to the regression problem belongs to a subspace generated by the feature maps at the data points received. Then, we formulate a LS based regression problem for each agent and find closed form expressions for the optimal solution. After the regression at the individual agents is complete, the agents transmit the function they have learned to a fusion center. Since the dependency from input to output is the same for both agents, the objective of the fusion center is to combine or fuse the functions and declare the fused function as the function learned by the system. We present an optimization based framework for fusion of the functions received at the fusion center.

For motivation, we consider a simple scenario where two agents with different onboard sensors are collecting input and output data using the sensors. The true data gets corrupted by noise of the sensors. Depending on the quality of the sensors, sensor type, etc. , the data received by the agents is most likely going to be different. The problem we consider has some similarities and some dissimilarities to horizontal FL. It is similar in the aspect that, the sample spaces of the

*Research supported by the Swedish Research Council (VR), Swedish Foundation for Strategic Research (SSF), and the Knut and Alice Wallenberg Foundation. The authors are with the Division of Decision and Control Systems, Royal Institute of Technology, KTH, Stockholm. Email: aneesh@kth.se, kallej@kth.se

Authorized licensed use limited to: KTH Royal Institute of Technology. Downloaded on October 24,2025 at 15:54:23 UTC from IEEE Xplore. Restrictions apply.

agents have minimal intersection and the feature space is the same for the agents. The formulation is different from horizontal FL in the aspect that the function spaces over which we optimize are more structured and fusion problem is based on an optimization problem rather than weighted averages. We note that this paper is not about FL; we only draw comparison to FL as it is a distributed learning approach that is being actively researched up on. The problem we consider is also different from the formulations studied in statistical learning theory (SLT). In [15], [16], the input and output data are random variables with the output being a function of the input. The joint distribution between the input and output random variables is unknown. In our formulation, the true input and output data are deterministic, while joint distribution between the received noisy data is known as joint distribution between the noise in the input and output data is known.

The outline of the paper is follows. In the next section, section II, we begin with a brief introduction to kernels and the spaces generated by them. We then discuss the noise model followed by the formulation of a general regression and LS regression problem with noisy data. In section III, we present the solutions to the problems formulated in section II. In section IV, we formulate the fusion problem as an optimization problem. In section V, we apply the methodology developed in sections III and IV to data generated from a transcendental function. We compare our solution to a solution obtained using methods from literature. In section VI, we conclude with some final remarks and discuss future work. Notation: we use superscript for the agent, subscript for samples and summation indicies. We represent vectors obtained by concatenating smaller vectors in boldface.

II. Problem Formulation

A. Background

We begin with a brief introduction to kernels and the function spaces generated by them. Let $\mathcal{X}$ be a nonempty subset of $\mathbb{R}^d, d \in \mathbb{N}$. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a function. Let $n \in \mathbb{N}$ and $\{x_1, \ldots, x_n\} \subset \mathcal{X}$. The $n \times n$ matrix, $\mathbf{K} := (K(x_i, x_j))_{ij}$, is called the *Gram (kernel)* matrix of K with respect to $x_1, \ldots, x_n$.

Definition 1 (Positive Definite kernel, [5])**.** A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which for all $n \in \mathbb{N}$ and for any set $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ gives rise to positive definite Gram matrix is called as positive definite kernel.

For a fixed x, the function $K(\cdot, x)$ is referred to as *feature map* at x. It is also denoted by $\Phi(x)$, i.e., $\Phi(x) = K(\cdot, x)$. Given a positive definite kernel $K : E \times E \to \mathbb{R}$, let H_0 be the linear space obtained from different linear combinations of the feature maps at different points in the domain:

$$H_0 = \{f(x) : f(x) = \sum_{i=1}^{n} \alpha_j K(x,x_j), \{\alpha_j\}_{j=1}^n \in \mathbb{R}, \{x_i\}_{j=1}^n \in \mathcal{X}, n \in \mathbb{N}\}$$

Given $f(x) = \sum_{j=1}^n \alpha_j K(x, x_j)$ and $g(x) = \sum_{k=1}^m \beta_k K(x, y_k)$ belonging to H_0, we define a bilinear form $\langle f, g\rangle$ on H_0 as:

$$\langle f, g\rangle = \sum_{j=1}^{n}\sum_{k=1}^{m} \alpha_j \beta_k K(y_k, x_j).$$

It can be verified that the inner product does not depend on the representation of the functions f and g. The norm induced by the inner product is $||f|| = \sqrt{\langle f, f\rangle}$. Let H be the space obtained by taking the completion of H_0 with respect to this norm. Then H is the RKHS generated by K.

B. Noise Model

The noisy data for agent i represented by $\big((x_1^i, y_1^i), \ldots, (x_m^i, y_m^i)\big)$, where x_j^i, y_j^i and m are input data, output data and number of samples respectively. The noise in the input and output of function are denoted by ε_j^i and η_j^i respectively, i.e,

$$x_j^i = \bar{x}_j^i + \varepsilon_j^i,\ y_j^i = \bar{y}_j^i + \eta_j^i,$$

where $(\bar{x}_j^i, \bar{y}_j^i)$ are the true values of the data. We assume that the noise variables ε_j^i and η_j^i take values in the set $\hat{\mathcal{X}}^i \subset \mathbb{R}^d$ and $\hat{\mathcal{Y}}^i \subset \mathbb{R}$ respectively. We assume that true data points $(\bar{x}_j^i, \bar{y}_j^i)$ take values in $\bar{\mathcal{X}} \subset \mathbb{R}^d$ and $\mathbb{R}$ respectively. We denote the Minkowski sum of $\hat{\mathcal{X}} \oplus \bar{\mathcal{X}}$ as $\mathcal{X} \subset \mathbb{R}^d$, where $\hat{\mathcal{X}} = \hat{\mathcal{X}}^1 \bigcup \hat{\mathcal{X}}^2$. We assume that the true joint distribution of $(\varepsilon_j^i, \eta_k^i)$, for $j = 1, \ldots, m$ and $k = 1, \ldots, m$ is known (for a given i). Thus, we assume that only the local joint distributions of the noise are known. We consider the probability space at agent i. The sample space Ω^i is a set of $2m$-tuples, where the first m-tuple (of a $2m$-tuple) consists of the noise in $\{x_j^i\}_{j=1}^m$, while the second m-tuple consist of noise in $\{y_j^i\}_{j=1}^m$ and hence is equal to $\hat{\mathcal{X}^i}^m \times \hat{\mathcal{Y}^i}^m$. With the sample space we associate a sigma algebra, F^i. F^i could be set of all subsets of Ω^i, i.e. , the power set of Ω^i, when Ω^i is finite space or could be the sigma algebra generated by all open subsets of Ω^i (depending on the topology on Ω^i). The known joint distribution of noise at agent i is denoted by P^i while the expectation with respect to measure P^i is denoted as $\mathbb{E}_{P^i}$. Thus, the probability space for agent i is (Ω^i, F^i, P^i) and $\{\varepsilon_j^i, \eta_j^i\}_{j=1}^m$ are random variables on this space.

The difference between our formulation and the formulation of function learning problems in the SLT literature, [15], [16], is explained as follows. In the traditional formulation, the input data, $\{X_i\}$, are random variables drawn from a known distribution $P(\mathcal{X})$. The outputs, $\{Y_i\}$ are random variables, given by $Y_i = f(X_i)$, where the function f is unknown. Hence the joint distribution between X_i and Y_i is unknown. To formulate the learning problems, the conditional, $P(\mathcal{Y}|\mathcal{X})$, is assumed be fixed, *unknown* and belonging to a family of distributions, $\mathcal{P}(\mathcal{Y}|\mathcal{X})$. Using the joint distribution obtained from the marginal and the conditional distribution, the learning problems are formulated as expected risk minimization problems . Using an induction principle, [15], the expected risk is replaced by the empirical risk. In our formulation, the true input

Authorized licensed use limited to: KTH Royal Institute of Technology. Downloaded on October 24,2025 at 15:54:23 UTC from IEEE Xplore. Restrictions apply.

data $\{\bar{x}_i\}$ and true output data $\{\bar{y}_i\}$ are deterministic, and $\bar{y}_i = f(\bar{x}_i)$. Noise is added to both input and output, $x_i = \bar{x}_i + \varepsilon_i$ and $y_i = \bar{y}_i + \eta_i$. Thus $\{x_i\}$ and $\{y_i\}$ are random variables whose joint distribution depends on the the joint distribution between $\{\varepsilon_i\}$ and $\{\eta_i\}$ which is *known*. If $\{\varepsilon_i\}$ and $\{\eta_i\}$ are independent, then noisy data points $\{x_i\}$ and $\{y_i\}$ are independent, which is not possible in the first formulation.

C. General Regression Problem

Let $C^i : (\mathcal{X} \times \mathbb{R} \times \mathbb{R})^m \to \mathbb{R}$ be a arbitrary loss function and $\Psi^i : [0, \infty) \to \mathbb{R}$ be a strictly monotonic increasing function. For any function $f \in H$, we define the total cost as:

$$\mathcal{C}^i(f) = \mathbb{E}_{P^i}\Big[C^i\Big((x_1^i - \varepsilon_1^i, y_1^i - \eta_1^i, f(x_1^i - \varepsilon_1^i)), \ldots, (x_m^i - \varepsilon_m^i, y_m^i - \eta_m^i, f(x_m^i - \varepsilon_m^i))\Big)\Big] + \Psi^i(||f||_H)$$

We note that the cost function $\mathcal{C}^i(\cdot)$ depends on the data points, $(x_j^i, y_j^i)_{j=1}^m$ as well. That is, for a given function, f, if the data points change the cost is also going to change. But the data set is not something that we can control or optimize over. It is given and hence we do not indicate the same in the cost function arguments.

1) Regression Problem 1: Consider the subspace of functions, $\bar{H}^i = \sum_{j=1}^{j=m} \alpha_j^i \mathbb{E}_{P^i}[K(x, x_j^i - \varepsilon_j^i)]$, where $\mathbb{E}_{P^i}[K(x, x_j^i - \varepsilon_j^i)] = \int_{\Omega^i} K(x, x_j^i - \varepsilon_j^i(\omega)) dP^i(\omega)$, the integral is defined in the Bochner sense, [17] and $\{\alpha_j^i\}_{j=1}^m \subset \mathbb{R}$. When $\hat{\mathcal{X}}^i$ is a finite set, $\mathbb{E}_{P^i}[K(x, x_j^i - \varepsilon_j^i)] = \sum_{\omega \in \Omega^i} K(x, x_j^i - \varepsilon_j^i(\omega)) P^i(\varepsilon_j^i(\omega))$. We define the first regression problem as an optimization problem for each agent,

$$(P1)^i : \min_{f \in \bar{H}^i} \mathcal{C}^i(f).$$

The subspace of functions over which we optimize draws inspiration from the *representer theorem*, [18]. In the deterministic case, i.e., when there is no noise in the data, the solution for optimizing $\mathcal{C}^i(f)$(instead the expectation of the first term, the first term itself) is given by a linear combination of the feature maps at the data points x_j^i, i.e., $f^{i^*} = \sum_j \alpha_j^i \Phi(x_j^i)$. In the noisy case, we want to find the optimal function by replacing $\Phi(x_j^i)$, by $\mathbb{E}_{P^i}[\Phi(x_j^i - \varepsilon_j^i)]$.

2) Regression Problem 2: We consider the regression problem as an optimization problem over the entire space of functions, rather than the subspace(s) considered before:

$$(P2)^i : \min_{f \in H} \quad \mathcal{C}^i(f).$$

D. LS Regression with Regularization

We consider the cost function,

$$C^i(\{\alpha_l, \beta_l, \gamma_l\}_{l=1}^m) = \sum_{l=1}^{m} (\beta_l - \gamma_l)^2$$

and $\Psi^i(x) = \varrho^i x^2$. We note that regression problem reduces to the problem considered in the Gauss -Markov Theorem (nonrandom parameter estimation, [19]) when there is no noise in the input data and the matrix considered is the matrix generated from the kernel.

III. Solution

A. The Stochastic Representer Theorem

For agent i, let $\boldsymbol{\alpha}^i = [\alpha_1^i; \ldots; \alpha_m^i]$, where α_j^i are random variables in $L^\infty(\Omega^i, F^i, P^i)$. Let $\boldsymbol{\Phi}^{\mathbf{i}}(\mathbf{x} - \boldsymbol{\varepsilon}) = [\Phi(x_1^i - \varepsilon_1^i), \ldots, \Phi(x_m^i - \varepsilon_m^i)]$. Consider the following subspace of the RKHS H,

$$\mathfrak{M}^i = \Big\{ f \in H : f = \mathbb{E}_{P^i}[{\boldsymbol{\alpha}^i}^T \boldsymbol{\Phi}^{\mathbf{i}}(\mathbf{x} - \boldsymbol{\varepsilon})], \; \boldsymbol{\alpha}^i \in \Big\{ L^\infty(\Omega^i, F^i, P^i) \Big\}^m \Big\},$$

and let $\bar{\mathfrak{M}}^i$ be its closure with respect to the norm topology.

Assumption 1. *We assume that $\mathcal{X}^i$ and $\mathcal{Y}^i$ are finite sets, i.e., the noise in input and output data take values in finite sets.*

Lemma III.1. $\Phi(x_j^i - \varepsilon_j^i(\omega)) \in \bar{\mathfrak{M}}^i, \forall \omega \in \Omega^i \;, \forall j = 1, \ldots, m.$

Proof. When Ω^i is finite and thus the joint distribution of the noise, P^i are probability mass functions, $\mathfrak{M}^i = \Big\{ f \in H : f = \sum_{\omega \in \Omega^i} \sum_{j=1}^m \alpha_j^i(\omega) K(x, x_j^i - \varepsilon_j^i(\omega)) P^i(\omega) \Big\}$. $\mathfrak{M}^i$ is generated by linear combinations of a finite subset of vectors in H. Thus, $\mathfrak{M}^i$ is a finite dimensional subspace of H and is closed, i.e., $\mathfrak{M}^i = \bar{\mathfrak{M}}^i$. Let $\omega^* \in \Omega^i$ and $j^* \in \{1, \ldots, m\}$. We define $\alpha_{j^*}^i(\omega^*) = \dfrac{1}{P^i(\omega^*)}$, $\alpha_{j^*}^i(\omega) = 0$ for all other ω, $\alpha_j^i(\omega) = 0$ for all other j and all ω. Expressing the same in a compact form, $\alpha_{j^*}^i = \dfrac{\chi_{\omega^*}(\omega)}{P^i(\omega^*)} = \dfrac{\chi_{\omega^*}(\omega)}{\mathbb{E}_{P^i}(\chi_{\omega^*})}$, where $\chi(\cdot)$ is the indicator function. Thus, $\boldsymbol{\alpha}^i = [0; \ldots; \alpha_{j^*}^i; \ldots; 0]$ and, $\mathbb{E}_{P^i}[{\boldsymbol{\alpha}^i}^T \boldsymbol{\Phi}^{\mathbf{i}}(\mathbf{x} - \boldsymbol{\varepsilon})] = \Phi(x_{j^*}^i - \varepsilon_{j^*}^i(\omega^*)) \in \bar{\mathfrak{M}}^i$. □

Theorem III.2. *The solution to problem $(P2)^i$ belongs to the closed subspace $\bar{\mathfrak{M}}^i$.*

Proof. Since $\bar{\mathfrak{M}}^i$ is closed subspace, $H = \bar{\mathfrak{M}}^i \oplus \bar{\mathfrak{M}}^{i^\perp}$. Any vector $f \in H$ can be expressed as $f^i + \bar{f}^i$ where $f^i \in \bar{\mathfrak{M}}^i$ and $\bar{f}^i \in \bar{\mathfrak{M}}^{i^\perp}$. We note that the cost function, $\mathcal{C}^i(f)$ depends on the function value at $x_j^i - \varepsilon_j^i(\omega)$, $f(x_j^i - \varepsilon_j^i(\omega))$, $\omega \in \Omega^i$ and the norm of the function , $||f||_H$. We note that $x_j^i - \varepsilon_j^i(\omega)$ is not a random variable but a member of $\mathcal{X}$. By the reproducing property of the kernel, $f(x_j^i - \varepsilon_j^i(\omega)) = \langle f(\cdot), K(\cdot, x_j^i - \varepsilon_j^i(\omega)) \rangle_H = \langle f^i(\cdot) + \bar{f}^i(\cdot), K(\cdot, x_j^i - \varepsilon_j^i(\omega)) \rangle_H$. By Lemma III.1, $K(\cdot, x_j^i - \varepsilon_j^i(\omega)) = \Phi(x_j^i - \varepsilon_j^i(\omega)) \in \bar{\mathfrak{M}}^i$ and, is thus orthogonal to $\bar{f}^i$. Thus $\langle \bar{f}^i(\cdot), K(\cdot, x_j^i - \varepsilon_j^i(\omega)) \rangle_H = 0$ and $f(x_j^i - \varepsilon_j^i(\omega)) = f^i(x_j^i - \varepsilon_j^i(\omega))$. Hence the first term in the cost function $\mathcal{C}^i(f)$, does not depend on $\bar{f}^i$. $||f||_H^2 = ||f^i||_H^2 + ||\bar{f}^i||_H^2$, which is minimized when $\bar{f}^i$ is equal to zero. Thus the optimizer of $(P2)^i$, f^* belongs to $\bar{\mathfrak{M}}^i$. □

This theorem can be viewed as an analogue to the classic *representer theorem*. It is necessary for the $\{\alpha_j^i\}$ to be random variables, so that the feature maps, $\{K(x, x_j^i - \varepsilon_j^i(\omega))\}_{\omega \in \Omega^i}$, are included in the subspace, $\bar{\mathfrak{M}}^i$, which is a key requirement to prove the theorem.

Authorized licensed use limited to: KTH Royal Institute of Technology. Downloaded on October 24,2025 at 15:54:23 UTC from IEEE Xplore. Restrictions apply.

B. LS Regression with Regularization

Let $\bar{\mathbf{K}}^i(\zeta) = \Big(\mathbb{E}_{P^i}\Big[K(x_j^i - \varepsilon_j^i(\zeta), x_k^i - \varepsilon_k^i)\Big]\Big)_{jk}$, where the expectation is only with respect to the second argument and $\mathbb{E}[\mathbf{K}^i] = \Big(\mathbb{E}_{P^i\times P^i}\Big[K(x_j^i - \varepsilon_j^i, x_k^i - \varepsilon_k^i)\Big]\Big)_{jk}$, where the expectation is with respect to the product measure of P^i with itself. We note that $\mathbb{E}_{P^i}[\bar{\mathbf{K}}^i] = \mathbb{E}[\mathbf{K}^i]$. Let $\mathbf{y}^i - \boldsymbol{\eta}^i = [y_1^i - \eta_1^i, \ldots, y_m^i - \eta_m^i]$. The LS regression problem is to minimize,

$$\mathcal{C}^i(f) = \mathbb{E}_{P_i}\Big[\sum_{j=1}^{m}(y_j^i - \eta_j^i - f(x_j^i - \varepsilon_j^i))^2\Big] + \varrho^i||f||_H^2.$$

Proposition III.3. *The optimal solution to $(P1)^i$ is given by $\boldsymbol{\alpha}^{i^*^T}\mathbb{E}_{P^i}[\boldsymbol{\Phi}^{\mathbf{i}}(\mathbf{x}-\boldsymbol{\varepsilon})]$, where*

$$\boldsymbol{\alpha}^{i^*} = \Big[\Big[\mathbb{E}_{P^i}[\mathbf{K}^{i^T}\mathbf{K}^i] + \varrho^i\mathbb{E}[\mathbf{K}^i]\Big]^T\Big]^{-1}\Big[\mathbb{E}_{P^i}[\bar{\mathbf{K}}^{i^T}(\mathbf{y}^i - \boldsymbol{\eta}^i)]\Big]. \tag{1}$$

Proof. The optimization is over the space of functions, $\bar{H}^i = \{f \in H : f = \boldsymbol{\alpha}^T\mathbb{E}_{P^i}[\boldsymbol{\Phi}^{\mathbf{i}}(\mathbf{x}-\boldsymbol{\varepsilon})], \boldsymbol{\alpha} \in \mathbb{R}^m\}$. For any $f \in \bar{H}^i$, by the reproducing property,

$$\begin{aligned}
&f(x_j^i - \varepsilon_j^i(\zeta)) = \langle f(\cdot), \Phi(x_j^i - \varepsilon_j^i(\zeta))\rangle\\
&= \langle\sum_k \alpha_k^i\mathbb{E}_{P^i}[K(x, x_k^i - \varepsilon_k^i)], \Phi(x_j^i - \varepsilon_j^i(\zeta))\rangle\\
&\overset{(a)}{=} \sum_k \alpha_k^i \sum_{\omega\in\Omega^i}\langle K(x, x_k^i - \varepsilon_k^i(\omega)), \Phi(x_j^i - \varepsilon_j^i(\zeta))\rangle P^i(\omega)\\
&= \sum_k \alpha_k^i\mathbb{E}_{P^i}[K(x_j^i - \varepsilon_j^i(\zeta), x_k^i - \varepsilon_k^i)].\\
&\Rightarrow [f(x_1^i - \varepsilon_1^i(\zeta)), \ldots, f(x_m^i - \varepsilon_m^i(\zeta))] = \bar{\mathbf{K}}^i(\zeta)\boldsymbol{\alpha}^i.
\end{aligned}$$

Equality $\overset{(a)}{=}$ follows from linearity of the inner product in the first argument. The norm of f, $||f||_H^2$, is calculated as in equation set (2). Equality $\overset{(b)}{=}$ follows from linearity of the inner product in both its arguments. Equality $\overset{(c)}{=}$ follows from the reproducing property of the kernel and the construction of the product space given two probability spaces. Since f is parameterised by $\boldsymbol{\alpha}$, the cost function as function of $\boldsymbol{\alpha}$, $\mathcal{C}^i(\boldsymbol{\alpha}) := \mathcal{C}^i(f)$, can be expressed as $\mathbb{E}_{P^i}[((\mathbf{y}^i - \boldsymbol{\eta}^i) - \bar{\mathbf{K}}^i\boldsymbol{\alpha}^i)^T((\mathbf{y}^i - \boldsymbol{\eta}^i) - \bar{\mathbf{K}}^i\boldsymbol{\alpha}^i)] + \varrho^i\boldsymbol{\alpha}^{i^T}\mathbb{E}[\mathbf{K}^i]\boldsymbol{\alpha}^i$. From the first order necessary conditions of optimality, $\boldsymbol{\alpha}^{i^*}$ is found by setting the gradient of $\mathcal{C}^i(\boldsymbol{\alpha})$ to zero, $\nabla_\alpha\mathcal{C}^i(\boldsymbol{\alpha}^{i^*}) = 0$.

$$\begin{aligned}
\nabla_{\boldsymbol{\alpha}}\mathcal{C}^i(\boldsymbol{\alpha}^{i^*})) = -2\mathbb{E}_{P^i}[\bar{\mathbf{K}}^{i^T}(\mathbf{y}^i - \boldsymbol{\eta}^i)] + 2[\mathbb{E}_{P^i}[\mathbf{K}^{i^T}\mathbf{K}^i]^T +\\
\varrho^i\mathbb{E}[\mathbf{K}^i]^T]\alpha^{i^*} = 0.
\end{aligned}$$

Thus, $\boldsymbol{\alpha}^{i^*}$ satisfies equation (1). □

Proposition III.4. *The optimal solution to $(P2)^i$ is given by $\mathbb{E}_{P^i}\Big[\boldsymbol{\beta}^{i^*^T}\boldsymbol{\Phi}^{\mathbf{i}}(\mathbf{x}-\boldsymbol{\varepsilon})\Big]$, where*

$$\boldsymbol{\beta}^{i^*} = \Big[\hat{\mathbf{K}}^{i^T}\hat{\mathbf{K}}^i + \varrho^i\hat{\mathbf{K}}^i\Big]^{-1}\Big[\hat{\mathbf{K}}^i(\mathbf{y}^i - \boldsymbol{\eta}^i)\Big], \tag{3}$$

and $\hat{\mathbf{K}}^i(\zeta,\omega) = \Big(K(x_j^i - \varepsilon_j^i(\zeta), x_k^i - \varepsilon_k^i(\omega))\Big)_{jk}$ is a random matrix.

Proof. From Theorem III.2, the optimal solution belongs to $\bar{\mathfrak{M}}^i$. This subspace is parameterised by $\boldsymbol{\beta} \in \Big\{L^\infty(\Omega^i, F^i, P^i)\Big\}^m$. Following the steps as in the proof of the previous proposition, proposition III.3, we can prove that the cost functional, $\mathcal{C}^i(\boldsymbol{\beta}) := \mathcal{C}^i(f)$, is equal to,

$$\begin{aligned}
\mathcal{C}^i(\boldsymbol{\beta}) =&\mathbb{E}_{P^i}\Big[(\mathbf{y}^i - \boldsymbol{\eta}^i)^T(\mathbf{y}^i - \boldsymbol{\eta}^i) - 2(\mathbf{y}^i - \boldsymbol{\eta}^i)^T\mathbb{E}_{P^i}[\hat{\mathbf{K}}^i\boldsymbol{\beta}]\\
&+\mathbb{E}_{P^i\times P^i}[\boldsymbol{\beta}^T\hat{\mathbf{K}}^{i^T}\hat{\mathbf{K}}^i\boldsymbol{\beta}]\Big] + \varrho^i\mathbb{E}_{P^i\times P^i}[\boldsymbol{\beta}^T\hat{\mathbf{K}}^i\boldsymbol{\beta}]
\end{aligned}$$

Since our cost functional is defined over a space of functions, rather than a vector in $\mathbb{R}^m$, we find the *Gaîeaux* derivative of the functional. The *Gaîeaux* derivative of $\mathcal{C}^i(\cdot)$ at $\boldsymbol{\beta}$, is a bounded linear operator $T_{\mathfrak{G},\boldsymbol{\beta}}$, such that

$$\lim_{a\to 0}\frac{\mathcal{C}^i(\boldsymbol{\beta} + a\boldsymbol{\gamma}) - \mathcal{C}^i(\boldsymbol{\beta})}{a} = T_{\mathfrak{G},\boldsymbol{\beta}}(\boldsymbol{\gamma}), \forall\boldsymbol{\gamma}\in\Big\{L^\infty(\Omega^i, F^i, P^i)\Big\}^m.$$

After expanding the terms and taking limits, we can find that,

$$\begin{aligned}
T_{\mathfrak{G},\boldsymbol{\beta}}(\boldsymbol{\gamma}) = \mathbb{E}_{P^i}\Big[- 2(\mathbf{y}^i - \boldsymbol{\eta}^i)^T\mathbb{E}_{P^i}[\hat{\mathbf{K}}^i\boldsymbol{\gamma}] +\\
2\mathbb{E}_{P^i\times P^i}[\boldsymbol{\beta}^T\hat{\mathbf{K}}^{i^T}\hat{\mathbf{K}}^i\boldsymbol{\gamma}]\Big] + 2\varrho^i\mathbb{E}_{P^i\times P^i}[\boldsymbol{\beta}^T\hat{\mathbf{K}}^i\boldsymbol{\gamma}].
\end{aligned}$$

It can be verified that the *Fréchet* derivative of $\mathcal{C}^i(\cdot)$ is equal to $T_{\mathfrak{G},\boldsymbol{\beta}}$. The first order necessary condition for optimality, is that $T_{\mathfrak{G},\boldsymbol{\beta}}(\boldsymbol{\gamma}) = 0\ \forall\gamma$. Regrouping the terms and the expectations,

$$\begin{aligned}
&\mathbb{E}_{P^i}\Big[\mathbb{E}_{P^i\times P^i}\Big[- 2(\mathbf{y}^i - \boldsymbol{\eta}^i)^T\hat{\mathbf{K}}^i\boldsymbol{\gamma} + 2\boldsymbol{\beta}^T\hat{\mathbf{K}}^{i^T}\hat{\mathbf{K}}^i\boldsymbol{\gamma} +\\
&\qquad 2\varrho^i\boldsymbol{\beta}^T\hat{\mathbf{K}}^i\boldsymbol{\gamma}\Big]\Big] = 0\ \forall\gamma.\\
&\iff \hat{\mathbf{K}}^{i^T}\hat{\mathbf{K}}^i\boldsymbol{\beta} + \varrho^i\hat{\mathbf{K}}^{i^T}\boldsymbol{\beta} - \hat{\mathbf{K}}^{i^T}(\mathbf{y}^i - \boldsymbol{\eta}^i) = 0\ P^i\ \textbf{a.s.}
\end{aligned}$$

We denote the above expression by $\Gamma(\boldsymbol{\beta})$. The '$\Longleftarrow$' part is straightforward. For the '$\Longrightarrow$' part, we can argue by contradiction. Suppose $\Gamma(\boldsymbol{\beta})$ is not zero with probability 1. Since $\Gamma(\boldsymbol{\beta}(\omega)) \in \mathbb{R}^m$, there is atleast one index, j, for which $(\Gamma(\boldsymbol{\beta}))_j \neq 0$ on set a $A_j, P^i(A_j) > 0$. Let $\bar{A}_j = \{\omega \in A_j : (\Gamma(\beta(\omega)))_j > 0\}$. Define $\gamma_j(\omega) = 1,\ \omega \in \bar{A}_j$ and $\gamma_j(\omega)) = -1,\ \omega \in A_j \sim \bar{A}_j$, i.e. , $\gamma_j(\omega) = \chi_{\bar{A}_j}(\omega) - \chi_{A_j\sim\bar{A}_j}(\omega)$. All other components of $\boldsymbol{\gamma}$ are set to zero. For this $\boldsymbol{\gamma}$, $T_{\mathfrak{G},\boldsymbol{\beta}}(\boldsymbol{\gamma}) = \mathbb{E}_{P^i}[(\Gamma(\boldsymbol{\beta}))^T\boldsymbol{\gamma}] > 0$, which is a contradiction. Thus, $\boldsymbol{\beta}^{i^*}$ satisfies equation 3. Since our cost functional is strictly convex, the minimizer is unique. Hence, $\boldsymbol{\beta}^{i^*} \in \mathfrak{M}^i \subset \bar{\mathfrak{M}}^i$ solves $(P2)^i$. □

IV. Fusion Problem

We consider a fusion center that receives function f^i from the agent i. The goal of the fusion center is to fuse the functions received, taking into account the problem under consideration, i.e., to find a mapping from input to output data. Some possible approaches include: (a) rule based: if the cost incurred by agent 1 is greater than the cost incurred agent 2 (or vice-versa), then optimal function is $f^2(f^1)$, (b) weighted average: $f^* = \dfrac{c^1f^2 + c^2f^1}{c^1 + c^2}$, where c^1 and c^2 are the local costs incurred by the agents after solving the learning problem. Since these approaches are adhoc, we formulate the fusion problem as optimization problem.

Authorized licensed use limited to: KTH Royal Institute of Technology. Downloaded on October 24,2025 at 15:54:23 UTC from IEEE Xplore. Restrictions apply.

$$||f||_H^2 = \langle \sum_k \alpha_k^i \mathbb{E}_{P^i}[K(x, x_k^i - \varepsilon_k^i)], \sum_l \alpha_l^i \mathbb{E}_{P^i}[K(x, x_l^i - \varepsilon_l^i)] \rangle = \sum_{k,l} \alpha_k^i \alpha_l^i \langle \mathbb{E}_{P^i}[K(x, x_k^i - \varepsilon_k^i)], \mathbb{E}_{P^i}[K(x, x_l^i - \varepsilon_l^i)] \rangle$$

$$= \sum_{k,l} \alpha_k^i \alpha_l^i \langle \sum_{\zeta \in \Omega^i} K(x, x_k^i - \varepsilon_k^i(\zeta)) P^i(\zeta), \sum_{\omega \in \Omega^i} K(x, x_l^i - \varepsilon_l^i(\omega)) P^i(\omega) \rangle \overset{(b)}{=} \sum_{k,l} \alpha_k^i \alpha_l^i \sum_{\zeta \in \Omega^i} \sum_{\omega \in \Omega^i} \langle K(x, x_k^i - \varepsilon_k^i(\zeta)),$$

$$K(x, x_l^i - \varepsilon_l^i(\omega)) \rangle P^i(\zeta) P^i(\omega) \overset{(c)}{=} \sum_{k,l} \alpha_k^i \alpha_l^i \sum_{\omega \in \Omega^i \times \Omega^i} K(x_l^i - \varepsilon_l^i(\omega), x_k^i - \varepsilon_k^i(\omega)) (P^i \times P^i)(\omega) = {\boldsymbol{\alpha}^i}^T \mathbb{E}[\mathbf{K}^i] \boldsymbol{\alpha}^i. \quad (2)$$

The fusion center considers a set of functions $\{\mathfrak{b} = \{b_k\}_{k \geq 1} \subset H\}$ which span H to define a dissimilarity measure between f^1 and f^2 as:

$$d_{\mathfrak{b}}(f, g) = \sum_k \langle f - g, b_k \rangle_H^2.$$

We note that $d_{\mathfrak{b}}(\cdot, \cdot)$ need not be a metric on H; it depends on the choice of the set $\mathfrak{b}$. When the set $\mathfrak{b}$ is a set of orthonormal basis vectors, $d_{\mathfrak{b}}(f, g) = ||f - g||_H^2$, then it is indeed a metric. The objective of the fusion center is to find a linear combination of f^1 and f^2, f^*, such that the dissimilarity between f^1, f^* and f^2, f^* is minimized. We consider a simple optimization problem for the same:

$$(P4) \min_{a,b \in \mathbb{R}} d_{\mathfrak{b}}(af^1 + bf^2, f^1) + d_{\mathfrak{b}}(af^1 + bf^2, f^2) + \lambda ||af^1 + bf^2||_H^2$$

The above convex optimization problem is to be solved computationally. The motivation to consider a set of functions $\mathfrak{b}$, is to define a dissimilarity measure that accounts for functional values at specific points (through features maps $\{\Phi(\cdot)\}$), and the function norm (through orthonormal basis vectors).

V. Example

In this section, we apply the solution and fusion method developed in sections III and IV to a specific example. The example is described as follows. True input and output data are generated using a trancendental function. The function is generated using monomial functions and trigonometric functions while coefficients of the terms have been chosen randomly. Using the true function, the true input and output data are generated by considering 100 data points uniformly spaced on the input domain $[-3, 3]$. Noisy data is generated by adding noise to both the input and the output. The distribution of the noise data considered for the agents in mentioned in table I. We note that the noise in the input is considered to be *independent* of the noise in the output, hence the joint distribution of the noise in the input and output is the product distribution. The noise distribution for the agents has been chosen such that agent 1 has "more" noise in the input than agent 2, while agent 2 has "more" noise in the output than agent 1. The noisy data for agent 1 and agent 2 are plotted in figures 1 and 2 respectively.

Given noisy data, for both agents, first we consider least squares based curve fitting. In this approach, we consider the true output data to be a polynomial function of the input data, i.e. , $f(x) = a_0 + a_1 x + \ldots + a_n x^n$. The noisy data received is assumed to be of the form $(x, f(x) + \varepsilon)$, where ε is noise in the received output data. No noise is considered in the input data. In the least squares based polynomial regression, the objective is to solve the optimization problem $\min_{a_0, \ldots, a_n} \mathbb{E}[\sum_i (y_i - f(x_i))^2]$, where the expectation is with respect to the noise distribution. The optimization problems are solved for polynomials of different degree. The "best" fit functions learned by the agents are plotted in figure 1 and 2 respectively in red. Next, we consider the functions learned by the agents by solving problem $(P1^i)$, II-C1. For both the agents, we consider a polynomial kernel, $K(x, y) = (x \times y)^d$. Given the noisy data, the functions learned by the agents by solving $(P1)^i$ have been plotted in figures 1 and 2 respectively.

Both approaches are tuned by varying the parameters appropriately. For $(P1)^i$, ϱ^i values where tuned, while for the regression, polynomials of different order are optimized over. It can be noted that the function learned from regression is less robust, i.e. , it has been overfit. At the fusion center, to define and solve $P4$, we consider $\mathfrak{b}$ to contain only functions of the form $K(\cdot, x_i)$, where x_i's were chosen as the points where f^1 and f^2 differed the most. These x_i's were obtained by inspection. In figure 3, we plot the average of the functions learned through curve fitting by the agents, the function obtained upon solving the fusion optimization problem $(P4)$ and the true function from which noisy data was generated. The true function has only been plotted for comparison purposes. In practice, the true function is not available. We note that our approach does not overfit and has better prediction capabilities for new data. We observe that there is lesser variation in functions learned by agents in our approach than the curve fitting approach, i.e., the functions learned by the agents by solving $(P1)^i$ are similar even though the noise distributions are different. Thus, our approach can be viewed as noise rejection (through expectation) followed by regression.

VI. Conclusion and Future Work

Hence, we considered the problem of learning of functions from (input, output) data corrupted by noise of known distribution. Regression problems were formulated over a RKHS generated by a kernel and a subspace of this function space, and their solutions were compared. The fusion problem resulted in a linear combination of the functions received at the fusion center, where the coefficients

Authorized licensed use limited to: KTH Royal Institute of Technology. Downloaded on October 24,2025 at 15:54:23 UTC from IEEE Xplore. Restrictions apply.

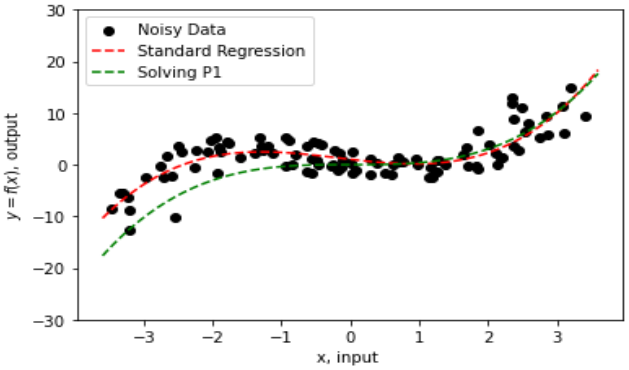

Fig. 1. Data for agent 1 and functions learned by agent 1

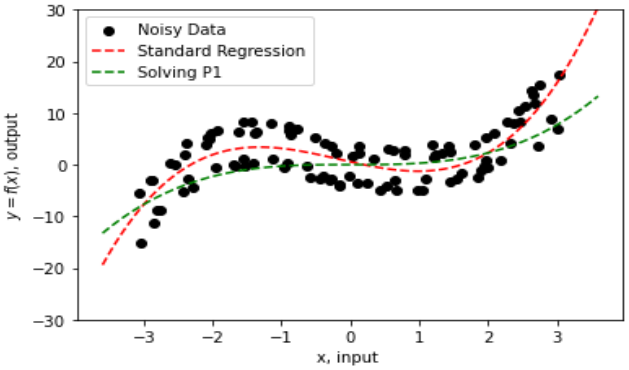

Fig. 2. Data for agent 2 and functions learned by agent 2

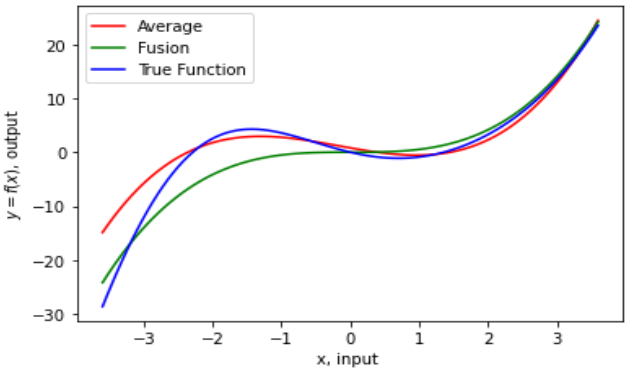

Fig. 3. At fusion center: average, fusion by optimization and true function

For agent 1:

Noise in x	-0.6	-0.4	-0.2	0.2	0.4	0.6
pmf	0.2	0.15	0.15	0.18	0.15	0.17
Noise in y	-1.5	-1.0	-0.5	0.5	1.0	1.5
pmf	0.10	0.10	0.30	0.30	0.10	0.10

For agent 2:

Noise in x	-0.2	-0.12	-0.04	0.04	0.12	0.2
pmf	0.1	0.20	0.20	0.20	0.20	0.10
Noise in y	-4	-3.0	-2.0	2.0	3.0	4.0
pmf	0.20	0.20	0.10	0.10	0.20	0.20

TABLE I
PROBABILITY MASS FUNCTIONS OF THE NOISE DISTRIBUTION

were found by solving an optimization problem. As future work, we are interested in studying regression problems where the agents learn in function spaces generated by different kernels. The mathematical structure of the fusion space, the representation of functions learned by the agents in the fusion space need to be investigated. Extensions to scenarios where more than 2 agents collect data can also be considered. A rigorous framework for the fusion problem is to be developed. The overall objective would be to develop a collaborative iterative learning scheme between agents through the fusion center. The impact of this learning scheme on decision-making and control is to be studied.

REFERENCES

[1] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He, "A survey on federated learning systems: vision, hype and reality for data privacy and protection," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[2] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[3] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, and H. V. Poor, "Federated learning for internet of things: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1622–1658, 2021.

[4] Y. Xianjia, J. P. Queralta, J. Heikkonen, and T. Westerlund, "Federated learning in robotic and autonomous systems," *Procedia Computer Science*, vol. 191, pp. 135–142, 2021.

[5] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *The Annals of Statistics*, vol. 36, no. 3, pp. 1171–1220, 2008.

[6] F. Pérez-Cruz and O. Bousquet, "Kernel methods and their potential use in signal processing," *IEEE Signal Processing Magazine*, vol. 21, no. 3, pp. 57–65, 2004.

[7] Y. Altun, T. Hofmann, and A. J. Smola, "Gaussian process classification for segmenting and annotating sequences," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 4.

[8] W. Chen, H. Shahabi, A. Shirzadi, H. Hong, A. Akgun, Y. Tian, J. Liu, A. Zhu, S. Li *et al.*, "Novel hybrid artificial intelligence approach of bivariate statistical-methods-based kernel logistic regression classifier for landslide susceptibility modeling," *Bulletin of Engineering Geology and the Environment*, vol. 78, no. 6, pp. 4397–4419, 2019.

[9] W. Chen, Y. Li, P. Tsangaratos, H. Shahabi, I. Ilia, W. Xue, and H. Bian, "Groundwater spring potential mapping using artificial intelligence approach based on kernel logistic regression, random forest, and alternating decision tree models," *Applied Sciences*, vol. 10, no. 2, p. 425, 2020.

[10] N. Subrahmanya and Y. C. Shin, "Sparse multiple kernel learning for signal processing applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 5, pp. 788–798, 2009.

[11] G. Ding, Q. Wu, Y.-D. Yao, J. Wang, and Y. Chen, "Kernel-based learning for statistical signal processing in cognitive radio networks: Theoretical foundations, example applications, and future directions," *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 126–136, 2013.

[12] D. Romero, M. Ma, and G. B. Giannakis, "Kernel-based reconstruction of graph signals," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 764–778, 2016.

[13] Q. Guo, "Distributed semi-supervised regression learning with coefficient regularization," *Results in Mathematics*, vol. 77, no. 2, p. 63, 2022.

[14] E. Dobriban and Y. Sheng, "Distributed linear regression by averaging," *Annals of Statistics*, vol. 49, no. 2, pp. 918–943, 2021.

[15] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999.

[16] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 1999.

[17] J. Mikusiński, "Chapter iii: The bochner integral," in *The Bochner Integral*. Springer, 1978, pp. 15–22.

[18] B. Schölkopf, A. J. Smola, F. Bach *et al.*, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[19] H. V. Poor, *An introduction to signal detection and estimation*. Springer Science & Business Media, 1998.

Authorized licensed use limited to: KTH Royal Institute of Technology. Downloaded on October 24,2025 at 15:54:23 UTC from IEEE Xplore. Restrictions apply.