Energy-Economic Multiple Incremental/Decremental Kernel Ridge Regression for Green Clouds

Bo-Wei Chen

Abstract—This study presents an energy-economic approach for incremental/decremental learning based on kernel ridge regression, a frequently used regressor on clouds. To avoid reanalyzing the entire dataset when data change, the proposed mechanism supports incremental/decremental processing for both single and multiple samples (i.e., batch processing). Moreover, incremental/decremental analyses in empirical and intrinsic space are also introduced to handle with data matrices with a large number of samples or feature dimensions. At the end of this study, we further the proposed mechanism to statistical Kernelized Bayesian Regression, so that incremental/decremental analyses become applicable.

Index Terms—Multiple incremental/decremental analysis, kernel ridge regression, kernelized Bayesian regression, Gaussian process, batch learning

I. INTRODUCTION

Kernel ridge regression (KRR) further advances ridge regression by converting feature space into hyperspace with the use of kernel functions, for example, polynomial functions and radial basis functions. As one of the popular classifiers and regressors, KRR has higher computational load than Support Vector Machines (SVMs) [2]. Although KRR has a closed-form solution, which involves the inverse of matrices, calculating these matrices degrades computational speed. Literature reviews [1] showed that the complexity of KRR [3] was as high as $O(N^3)$, whereas that of SVMs was $O(N_2)$, in which $N$ stands for the number of instances in data. Such a characteristic is a burden to cloud servers, which consume too much power for computation, not to mention incremental analysis.

Unlike typical learning, incremental analysis allows the system to add new training samples. Additionally, no retraining is required. Both single incremental and multiple incremental mechanisms are conducive to relief of computational load as earlier calculation results can be reserved for updating the new system in the future. Furthermore, if the size of data is far beyond the capability of one machine, especially when the memory space cannot accommodate the entire data at once, incremental analysis is a feasible solution.

In kernelized learning, many efforts have been devoted in incremental classifiers. Cauwenberghs and Poggio [4] established a milestone for kernelized learning as they discovered the equilibrium between old Lagrangian multipliers and new ones. A differential form was derived from the cost function of SVMs and the Karush–Kuhn–Tucker (KKT) [5] conditions. Such a differential form supported single incremental and decremental learning. The derivation was shown in a subsequent study [6]. A recursive procedure was introduced to update the matrix formed by the original support vectors and the kernel matrix when a single instance was changed. The authors also devised a strategy called “bookkeeping,” or the accounting strategy mentioned in [7], to determine the largest increment/decremental amount of existing Lagrangian multipliers. The model by Cauwenberghs and Poggio has inspired subsequent studies, for example [6], [7], and [8]. Laskov et al. [7] summarized the methodology developed in [4] by presenting a systematic analytical solution. Such a solution explicitly and clearly introduced the changes in Lagrangian multipliers with respect to three cases: Unbounded support vectors, bounded support vectors, and non-support vectors. Each vector corresponded to one Lagrangian multiplier. Furthermore, they also presented recursive matrix updates and matrix decomposition that were conducive to incremental/decremental matrix computation. Karasuyama and Takeuchi [8] advanced the approach proposed by [4] and developed a strategy for multiple incremental/decremental learning. They simplified the bookkeeping strategy mentioned in [4] by searching the shortest and easiest path when existing Lagrangian multipliers were changed. The definition of the path in their works represented the series changes of increment/decremental values for existing Lagrangian multipliers. The major difference between [4] [7] and [8] is that the latter did not move the existing samples from their original regions (i.e., within the margin, on the margin, and outside the margin) into the other regions. The latter simply moved the incremental samples.

For KRR, incremental and decremental solutions become
easier compared with those of SVMs because KRR has a closed-form solution. Recent works, such as [9], [10] and [1], were examples for single-instance incremental regressors. Based on kernel concepts, Engel et al. [9] developed a kernel recursive least squares algorithm, or incremental kernel regression. Their fundamental idea was equivalent to ordinary least squares (OLS) or linear least squares in statistics, but performed in hyperspace. The same algorithm was employed by Vaerenbergh et al. [10]. They furthered incremental kernel regression and integrated it into uncertainty analyses. In [1], rather than focusing on kernel regression, a recursive version of KRR was introduced. It used a single incremental mechanism to update weight vectors of the cost function. Moreover, a forgetting factor was integrated into the recursive form, where old and new training samples had different weights.

As a statistical version of linear regression, Bayesian Linear Regression concentrates on uncertainty instead of frequencies. Its principal of finding weight vectors in the cost function is still the same as that of linear regression. Either Bayesian Ridge Regression or Kernelized Bayesian Ridge Regression is a special case of Bayesian Linear Regression. When the entire model focuses on the uncertainty in Gaussian distributions, Kernelized Bayesian Regression is equivalent to Gaussian process. Incremental Kernelized Bayesian Regression becomes a challenge because it has to deal with the product of a series of inverse matrices in the exponential form. Such a product usually comes from the parameters of Gaussian distributions: Means and covariance. Quinonero-Candela and Winther proposed an incremental solution for updating the parameters of Gaussian distributions, where a simplified prior model was used. They devised an Expectation–Maximization (EM) algorithm for proximally updating new means and covariance.

The contributions of this study are listed as follows.

- The proposed method supports decremental learning rather than merely incremental processing. Decremental learning is conducive to stabilizing the system by removing unnecessary outliers.
- Both incremental and decremental learning are capable of handling multiple samples in a batch.
- The proper size of a batch for incremental and decremental learning in the intrinsic space is derived in this article. Large-scale of datasets can be divided into several batches, which can be accommodated by a machine, for processing.
- The proposed mechanism furthers the earlier version of incremental Gaussian process by introducing decremental mechanisms and batch learning.

The rest of this paper is organized as follows. Section II introduces the multiple incremental/decremental learning in the intrinsic space for KRR, whereas Section III then describes the details of the learning in the empirical space. Next, Section IV extends the proposed mechanism to Kernelized Bayesian Regression. Conclusions are finally drawn in Section V.

II. INCREMENTAL/DECREMENTAL KERNEL RIDGE REGRESSION IN INTRINSIC SPACE

KRR has two types of operation modes. One is intrinsic space, and the other is empirical space. After feature mapping by kernel functions \( \phi \), intrinsic-space computation yields favorable complexity if the number of data \( N \) is far larger than the feature dimension \( M \). Otherwise, empirical-space operations should be used.

Let \( \{(x_i,y_i)\}_{i=1,...,N} \) denote a pair of an \( M \)-dimensional feature vector \( x_i \) and its corresponding label \( y_i \), where \( i \) specifies the indices of \( N \) training samples. The objective of a linear regressor is to minimize the following cost function of least-square errors (LSEs).

\[
\min_{\mathbf{u}, b} E_{KRR}(\mathbf{u}, b) = \min_{\mathbf{u}, b} \left\{ \sum_{i=1}^{N} \left( \mathbf{u}^T \phi(x_i) + b - y_i \right)^2 + \rho \|\mathbf{u}\|^2 \right\} \tag{1}
\]

where \( E_{KRR} \) is the cost function, \( \mathbf{u} \) represents a \( J \)-by-1 weight vector, \( \phi(x_i) \) denotes the intrinsic-space feature vector of \( x_i \), \( b \) is a bias term, and \( \rho \) specifies the ridge parameter. Besides, \( T \) means the conjugate operator, and \( \|\| \|^2 \) calculates the norm-two distance. Notable, \( J \) is the degree of intrinsic space when feature vectors are transformed by a kernel function.

Equation (1) can be rewritten as a matrix form, i.e.,

\[
E_{KRR}(\mathbf{u}, b) = \|\Phi^T \mathbf{u} + b \mathbf{e} - \mathbf{y}\|^2 + \rho \|\mathbf{u}\|^2. \tag{2}
\]

Individually differentiating (16) with respect to \( \mathbf{u} \) and \( b \), and subsequently zeroing both equations gives

\[
\mathbf{u} = \left( \Phi \Phi^T + \rho \mathbf{I} \right)^{-1} \Phi (\mathbf{y} - b \mathbf{e}) \tag{3}
\]

and

\[
b = \frac{1}{N} \left( \mathbf{e} \mathbf{y}^T - \mathbf{e}^T \Phi^T \mathbf{u} \right) \tag{4}
\]

Notice that \( \mathbf{K} = \Phi^T \Phi \) instead of \( \Phi \Phi^T \) mentioned in (3). Unlike the solution to kernel regression, i.e., \( \mathbf{u} = (\Phi \Phi^T)^+ \Phi (\mathbf{y} - b \mathbf{e}) \) where \( \Phi \Phi^T \) could be singular, KRR avoids such a problem by adding a ridge term inside.

A. Single Incremental and Decremental Processes

For intrinsic space, single incremental and decremental processes are straightforward. Let \( \mathbf{S} \) denote \( \Phi \Phi^T + \rho \mathbf{I} \). During the incremental phase, given a new training sample \( (x_\ell, y_\ell) \), the update of (3) becomes

\[
\mathbf{u}[\ell+1] = \mathbf{S}[\ell+1]^{-1} \mathbf{y}[\ell+1] - b[\ell+1] e[\ell+1] \tag{5}
\]
\begin{equation}
b[\ell+1] = \\
\frac{1}{N+1}(e[\ell+1]y^T[\ell+1] - e^T[\ell+1]\Phi^T[\ell+1]u[\ell+1]) \tag{6}
\end{equation}

where

\begin{equation}
\begin{aligned}
S[\ell+1]^{-1} &= \left(S[\ell] + \phi(x_c)\phi(x_c)^T \right)^{-1} \\
\Phi[\ell+1] &= \left[ \Phi[\ell] \phi(x_c) \right] \\
y[\ell+1] &= \left[ y[\ell] \ y_c \right]
\end{aligned} \tag{7}
\end{equation}

In (5)–(7), \( \ell \) denotes the current state of the system. To save the previous computation of \( S^{-1} \), the Sherman-Morrison formula and Woodbury matrix identity [11] indicates that

\begin{equation}
\begin{aligned}
S[\ell+1]^{-1} &= \left(S[\ell] + \phi(x_c)\phi(x_c)^T \right)^{-1} \\
&= S[\ell]^{-1} - \frac{S[\ell]^{-1}\phi(x_c)\phi(x_c)^T S[\ell]^{-1}}{1 + \phi(x_c)^T S[\ell]^{-1}\phi(x_c)} \tag{8}
\end{aligned}
\end{equation}

Regarding the decremental phase, given an index \( r \) of a sample, where \( r \in \{1, \ldots, N\} \),

\begin{equation}
\begin{aligned}
S[\ell-1]^{-1} &= \left(S[\ell] - \phi(x_r)\phi(x_r)^T \right)^{-1} \\
&= S[\ell]^{-1} - \frac{S[\ell]^{-1}\phi(x_r)\phi(x_r)^T S[\ell]^{-1}}{1 - \phi(x_r)^T S[\ell]^{-1}\phi(x_r)} \tag{9}
\end{aligned}
\end{equation}

For the rest parts, removing the \( r \)th sample from simply generates the recursive forms.

B. Multiple Incremental and Decremental Processes

For multiple incremental and decremental processes, assume the system is about to add \(|C|\) new samples and remove \(|R|\) existing data. The operator \(|\cdot|\) denotes the size. Then, (8) and (9) become

\begin{equation}
\begin{aligned}
S[\ell+1]^{-1} &= \left(S[\ell] + \Phi_C\Phi_C^T \right)^{-1} \\
&= S[\ell]^{-1} - \frac{S[\ell]^{-1}\Phi_C\left(1 + \Phi_C^T S[\ell]^{-1} \Phi_C \right)^{-1}\Phi_C^T S[\ell]^{-1}}{1 - \Phi_C^T S[\ell]^{-1}\Phi_C} \tag{10}
\end{aligned}
\end{equation}

and

\begin{equation}
\begin{aligned}
S[\ell-1]^{-1} &= \left(S[\ell] - \Phi_R\Phi_R^T \right)^{-1} \\
&= S[\ell]^{-1} - \frac{S[\ell]^{-1}\Phi_R\left(1 - \Phi_R^T S[\ell]^{-1} \Phi_R \right)^{-1}\Phi_R^T S[\ell]^{-1}}{1 - \Phi_R^T S[\ell]^{-1}\Phi_R} \tag{11}
\end{aligned}
\end{equation}

To facilitate multiple incremental and decremental processes at once, combination of (10) and (11) is necessary. Let \( \Phi_N = [\Phi_C \mid \Phi_R] \) represent the concatenation of all the column vectors in \( \Phi_C \) and \( \Phi_R \). Also denote \( \Phi'_N = [\Phi_C \mid -\Phi_R]^T \) as the concatenation of all the column vectors in \( \Phi_C \) and \( -\Phi_R \). Therefore, combination of (10) and (11) becomes

\begin{equation}
\begin{aligned}
S[\ell+1]^{-1} &= \left(S[\ell] + \Phi_C\Phi_C^T - \Phi_R\Phi_R^T \right)^{-1} \\
&= S[\ell]^{-1} - \frac{S[\ell]^{-1}\Phi_C\left(1 + \Phi_C^T S[\ell]^{-1} \Phi_C \right)^{-1}\Phi_C^T S[\ell]^{-1}}{1 - \Phi_C^T S[\ell]^{-1}\Phi_C} \tag{12}
\end{aligned}
\end{equation}

The batch sizes of \( \Phi_C \) and \( \Phi_R \), i.e., \(|C|\) and \(|R|\), can be different. Notably, the left-hand side of the two equations needs \( O(J^3) \), whereas the inverse on the right-hand side requires \( O(|C|^3) \) for (10) and \( O(|R|^3) \) for (11), respectively [12]. To ensure performance, when the number of samples in a batch is smaller than the size of intrinsic-space features (i.e., \(|C| < J \) or \(|R| < J \)), the system should perform an update. This implies a suitable batch size for streaming data, where new samples are rapidly generated and accumulated.

III. INCREMENTAL/DECREMENTAL KERNEL RIDGE REGRESSION IN EMPIRICAL SPACE

According to the Learning Subspace Property in [1], the weight vector \( u \) has the following relation between \( \Phi \) and an unknown \( N \)-dimensional vector \( a \).

\begin{equation}
u = \Phi a. \tag{13}
\end{equation}

Combining (2) and (13) yields

\begin{equation}
E_{\text{KRR}}(a,b) = \|Ka + be - y\|^2 + \rho a^T Ka. \tag{14}
\end{equation}

Rearranging the equations after differentiating (14) with respect to \( a \) and \( b \) yields

\begin{equation}
a = \left(K^T + \rho I \right)^{-1}(y - be) \tag{15}
\end{equation}

and
\[ b = \frac{y^T (K + \rho I)^{-1} e}{e^T (K + \rho I)^{-1} e}. \]  

(16)

A. Single Incremental and Decremental Processes

Given a new training sample \((x_r, y_r)\), the incremental phase is listed as follows.

\[
(K[\ell+1]+\rho I[\ell+1])^{-1} = [K[\ell]+\rho I[\ell] + \eta_c K_{c,c} + \rho]^{-1}
\]

(17)

where "\(^\sim\)\) signifies all the training samples excluding the new one. For simplicity, let \(Q\) denote \(K + \rho I\) and \(Q_{c,c}\) represent \(K_{c,c} + \rho\). Then, (17) becomes

\[
Q^{-1}[\ell+1] = \begin{bmatrix} Q[\ell] & \eta_c \\ \eta_c^T & Q_{c,c} \end{bmatrix}^{-1}
\]

(18)

However, (17) does not save the computational load as the system calculates the inverse again. According to the Sherman-Morrison formula and Woodbury matrix identity [7, 11], the inverse in (18) can be decomposed to two states. One is the current state \(Q^{-1}[\ell]\), and the other is \(Q^{-1}[\ell+1]\), shown as follows.

\[
Q^{-1}[\ell+1] = \begin{bmatrix} Q[\ell] & \eta_c \\ \eta_c^T & Q_{c,c} \end{bmatrix}^{-1}
\]

(19)

where

\[
\begin{align*}
B_{c,c} &= -Q^{-1}[\ell]\eta_c \\
\varepsilon &= Q_{c,c} - \eta_c^T Q^{-1}[\ell]\eta_c
\end{align*}
\]

(20)

Therefore, the computation of the inverse in the previous state can be kept for the next state. The incremental forms of (15) and (16) respectively become

\[
a[\ell+1] = (K^T[\ell+1]+\rho I[\ell+1])^{-1}(y[\ell+1]-b[\ell+1]e[\ell+1])
\]

(21)

and

\[
b[\ell+1] = \frac{y^T[\ell+1]Q^{-1}[\ell+1]e[\ell+1]}{e^T[\ell+1]Q^{-1}[\ell+1]e[\ell+1]}
\]

(22)

For the decremental phase, given an index \(r\) of a sample that is to be removed, where \(r \in \{1, \ldots, N\}\), we can re-arrange the elements in \(Q^{-1}\), so that \(r\) lies at the bottom-right corner of \(Q^{-1}\). Let \(\Theta, \xi_r, \xi, \eta_r, \eta\), respectively specify the four elements of \(Q^{-1}\), shown in (23). Besides, \(\Theta\) is a matrix, \(\xi_r\) denotes a vector, and \(\theta_r\) represents a scalar. Then,

\[
Q^{-1}[\ell-1] = \begin{bmatrix} \Theta & \xi_r \\ \xi_r^T & \theta_r \end{bmatrix} = \begin{bmatrix} Q^{-1}[\ell] + z^{-1}B_{c,c}B_{c,c}^T z^{-1}B_{c,c} \\
- z^{-1}B_{c,c}^T z^{-1} 
\end{bmatrix}
\]

(23)

The lower part comes from (19). Comparing the four terms in the upper and lower parts [7] yields the following form.

\[
Q^{-1}[\ell-1] = \Theta - \frac{\xi}{\theta_r^T}.
\]

(24)

Substitution (24) into (21) and (22) generates decremental forms.

B. Multiple Incremental and Decremental Processes

Like Section II.B, also assume the system adds \(|C|\) new samples and removes \(|R|\) existing data. For batch incremental learning, (19) becomes

\[
Q^{-1}[\ell+1] = \begin{bmatrix} Q^{-1}[\ell] + B_{c,c}z^{-1}B_{c,c}^T & z^{-1}B_{c,c} \\
- z^{-1}B_{c,c}^T & z^{-1} 
\end{bmatrix}
\]

(25)

where \(Z\) and \(B\) are matrices computed based on \(|C|\) new samples.

For batch decremental learning, we use (26) to replace (24).

\[
Q^{-1}[\ell-1] = \Theta - \xi_R \theta_R^{-1} \xi_R^T
\]

(26)

where \(\xi_R\) and \(\theta_R\) computed based on \(|R|\) existing data.

IV. INCREMENTAL/DECREMENTAL KERNELIZED BAYESIAN REGRESSION

Unlike KRR that focuses on frequentist methodologies, where sufficient occurrences are observed, Bayesian Regression concentrates on uncertainty. Thus, statistical distributions are introduced in Kernelized Bayesian Regression (KBR). As Bayesian theory works for various distributions, this work uses Gaussian distributions as a case study for demonstrate incremental and decremental analysis. As mentioned earlier in the introduction, such Kernelized Bayesian
Regression is equivalent to Gaussian process.

Consider a regression model,

\[ y_i = u^T \phi(x_i) + b \]  \hspace{1cm} (27)

or in a matrix form,

\[ y = u^T \Phi + b \]

where \( b \sim N(\mu_b, \Sigma_b) \) and \( x_i \in X \sim N(\mu_x, \Sigma_x) \). Furthermore, for simplicity, assume \( \mu_b = 0 \). Also, assume \( x_i \) and \( b \) are independent. Thus, \( \mu_y = u^T \mu_x + \mu_b = u^T \Sigma_x u + \Sigma_b \) and \( \Sigma_y = u^T \Sigma_u + \Sigma_u \). Furthermore, \( \Sigma_{xy} = \Sigma_{ux} \).

### A. Training Stage

The training stage uses Bayesian inference to estimate the posterior distribution of \( u \), which consists of two parts. One is the likelihood probability, and the other is the prior probability.

\[
P(u \mid x, y) = \frac{P(y \mid u, x)P(u)}{P(y \mid x)} \propto P(y \mid u, x)P(u) \]  \hspace{1cm} (28)

The following steps establish the posterior distribution by computing likelihood and prior probabilities.

#### ■ Computation of the Likelihood Probability

As \( b \sim N(0, \sigma_b^2 I) \) and \( x_i \sim N(\mu_x, \Sigma_x) \), the Gaussian distribution of the likelihood \( P(y \mid u, x) \) is \( N(\mu_{y \mid u, x}, \Sigma_{y \mid u, x}) \sim N(u^T \phi(x), \sigma_b^2 I) \) based on the following conditional expectation and conditional covariance of a linear Gaussian system [12].

\[
\Sigma_{y \mid x} = \Sigma_y - \Sigma_{yx} \Sigma_x^{-1} \Sigma_{xy}
\]  \hspace{1cm} (29)

\[
\mu_{y \mid x} = \mu_y + \Sigma_{yx} \Sigma_x^{-1} (\phi(x) - \mu_x)
\]  \hspace{1cm} (30)

Accordingly,

\[
P_y(u \mid x, y) = \prod_{i=1}^{N} P(y_i \mid u, x_i)
\]

\[
= \frac{1}{\sqrt{2\pi \Sigma_{y \mid u, x}}} e^{-\frac{1}{2} \sum_i (y_i - \mu_{y \mid i, u, x}) \Sigma_{y \mid u, x}^{-1} (y_i - \mu_{y \mid i, u, x})^T}
\]

\[
= \frac{1}{\sqrt{2\pi \sigma_b^2 I}} e^{-\frac{1}{2} \text{tr} \left( (y - \mu_{y \mid u, x}) \Sigma_{y \mid u, x}^{-1} (y - \mu_{y \mid u, x})^T \right)}
\]

\[
= \frac{1}{\sqrt{2\pi \sigma_b^2 I}} e^{-\frac{1}{2} \text{tr} \left( (y - \mu^T \phi(x)) (\sigma_b^2 I)^{-1} (y - \mu^T \phi(x))^T \right)}
\]

\[
(31)
\]

#### ■ Computation of the Prior Probability

Theoretically, the prior probability can be any distribution. However, such selection would result in posterior distributions without analytical solutions. This benefits no computation. As the likelihood probability is a Gaussian distribution, we can use conjugate prior distributions to model the system for convenience of computation. When the generated posterior distribution and the selected prior distribution belong to the same class of distributions, such a prior distribution is a conjugate prior distribution [13]. There is a systematic analytical model for conjugate prior distribution [14].

To generate a Gaussian posterior distribution, we consider the following Gaussian prior distribution.

\[
P(u) \sim N(\mu_u, \Sigma_u)
\]  \hspace{1cm} (32)

#### ■ Computation of the Posterior Probability

Recall that the likelihood \( P(y \mid u, x) \) is \( N(\mu_{y \mid u, x} = u^T \phi(x), \Sigma_{y \mid u, x} = \sigma_b^2 I) \), and the prior probability \( P(u) \) is \( N(\mu_u, \Sigma_u) \). According to [13] and [14], the posterior distribution is Gaussian.

\[
P(u \mid x, y) \propto P(y \mid u, x)P(u)
\]

\[
\propto N(\mu_{y \mid u, x}, \Sigma_{y \mid u, x}) N(\mu_u, \Sigma_u)
\]

\[
\propto N(u^T \phi(x), \sigma_b^2 I) N(\mu_u, \Sigma_u)
\]

\[
= N(u_{\text{eff}, u, x}, \Sigma_{\text{eff}, u, x})
\]  \hspace{1cm} (33)

where
\[
\begin{align*}
\Sigma_{\hat{u}y,x} & = \Sigma_u^{-1} + \sigma_b^2 \Phi \Phi^T \\
\mu_{\hat{u}y,x} & = \Sigma_{\hat{u}y,x}^{-1} \Sigma_u \mu_u + \sigma_b^2 \Phi \phi(x^*)
\end{align*}
\] (34)

B. Predicting Stage

As the training stage already generates the posterior distribution of \(u\), we use the posterior predictive distribution \(P(y^*|x^*,u)\) to model the uncertainty of the output. The posterior predictive distribution can be decomposed into the marginal distribution of \(P(y^*|x^*,u)\) and \(P(u|x,y)\). Let \(y^*\) represent the predictive output of the model when a test sample \(x^*\) is input. Applying the error-propagation rule of Gaussian identities [15] to the marginal distribution yields the following form.

\[
P(y^*|x^*,u) = \int P(y^*|x^*,u) P(u|x,y) du
\]

\[
= \int \mathcal{N} (\hat{u}, \hat{\Sigma}) \mathcal{N} (\mu_{\hat{u}y,x} | \Sigma_{\hat{u}y,x}) du
\]

\[
= \int \mathcal{N} (y^* | \mu_{\hat{u}y,x} \phi(x^*), \hat{\Sigma}) \mathcal{N} (u | \mu_{\hat{u}y,x}, \Sigma_{\hat{u}y,x}) du
\]

\[
= \mathcal{N} (\mu^*, \Sigma^*)
\]

Notably, \(P(y^*|x^*,u)\) is set to a Gaussian distribution so that the predictive distribution becomes Gaussian due to the need for analytical solutions.

Accordingly,

\[
P(y^*|x^*,u) = \frac{1}{\sqrt{2\pi} |\hat{\Sigma}|} e^{-\frac{1}{2} \text{tr} \left( (y^* - \mu^*) (\Sigma^*)^{-1} (y^* - \mu^*)^T \right)}
\] (36)

where

\[
\Sigma^* = \hat{\Sigma} + \phi(x^*)^T (\Sigma_u^{-1} + \sigma_b^2 \Phi \Phi^T)^{-1} \phi(x^*)
\] (37)

and

\[
\mu^* = (\Sigma_u^{-1} \mu_u + \sigma_b^2 \Phi \phi(x^*))^T (\Sigma_u^{-1} + \sigma_b^2 \Phi \Phi^T)^{-1} \phi(x^*)
\] (38)

V. CONCLUSION

This work presents an efficient incremental/decremental mechanism for KRR and KBR to update the weight vector of the cost function without retraining. The proposed mechanism uses a recursive structure to replace the original computation of weight vectors. Such a structure allows the system to utilize earlier results to update the system when either single or multiple samples are changed. This mechanism is conducive to save computational load on cloud servers. Therefore, incremental/decremental learning becomes applicable. This study also suggests an appropriate batch size during multiple incremental/decremental analyses in intrinsic space. The Mathematical model shows that the size of each batch should smaller than the feature dimensional size. Otherwise, it saves no computation.

REFERENCES

[1] S.-Y. Kung, Kernel Methods and Machine Learning. Cambridge, UK: Cambridge University Press, Jun. 2014.
[2] B.-W. Chen, C.-Y. Chen, and J.-F. Wang, “Smart homecare surveillance system: Behavior identification based on state transition support vector machines and sound directivity pattern analysis,” IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 43, no. 6, pp. 1279–1289, Nov. 2013.
[3] S.-Y. Kung and P.-Y. Wu, “On efficient learning and classification kernel methods,” in Proc. 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012), Kyoto, Japan, 2012, Mar. 25–30, pp. 2065–2068.
[4] G. Cauwenberghs and T. Poggio, “Incremental and decremental support vector machine learning,” in Proc. 14th Annual Conf. Neural Information Processing System (NIPS 2000), Denver, Colorado, United States, 2000, Nov. 28–30, pp. 409–415.
[5] J. C. Platt, “Fast training of support vector machines using sequential minimal optimization,” in Advances in Kernel Methods: Support Vector Learning, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA, USA: MIT Press, 1999.
[6] C. P. Diehl and G. Cauwenberghs, “SVM incremental learning, adaptation and optimization,” in Proc. International Joint Conference on Neural Networks (IJCNN 2003), Portland, Oregon, 2003, Jul. 20–24, pp. 2685–2690.
[7] P. Laskov, C. Gehl, S. Krüger, and K.-R. Müller, “Incremental support vector learning: Analysis, implementation and applications,” Journal of Machine Learning Research, vol. 7, pp. 1909–1936, 2006.
[8] M. Karasuyama and I. Takeuchi, “Multiple incremental decremental learning of support vector machines,” IEEE Transactions on Neural Networks, vol. 21, no. 7, pp. 1048–1059, Jul. 2010.
[9] Y. Engel, S. Mannor, and R. Meir, “The kernel recursive least-squares algorithm,” IEEE Transactions on Signal Processing, vol. 52, no. 8, pp. 2275–2285, Aug. 2004.
[10] S. V. Vaerenbergh, M. Lázaro-Gredilla, and I. Santamaria, “Kernel recursive least-squares tracker for time-varying regression,” IEEE Transactions on Neural Networks and Learning Systems, vol. 23, no. 8, pp. 1313–1326, Aug. 2012.
[11] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, Numerical Recipes 3rd Edition: The Art of Scientific Computing, 3rd ed. Cambridge, UK: Cambridge University Press, 2007.
[12] K. P. Murphy, Machine Learning: A Probabilistic Perspective. Cambridge, MA, USA: MIT Press, 2012.
[13] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, Bayesian Data Analysis, 3rd. ed. Boca Raton, Florida, USA: CRC Press, 2013.
[14] H. Raiffa and R. Schlaifer, Applied Statistical Decision Theory. Hoboken, New Jersey, USA: Wiley, 2000.
[15] M. Toussaint, “Introduction to machine learning,” Department of Computer Science, University of Stuttgart, Stuttgart, Germany, Apr. 19, 2016.