

# Accurate, Fast and Scalable Kernel Ridge Regression on Parallel and Distributed Systems

Yang You, James Demmel  
UC Berkeley  
{youyang,demmel}@cs.berkeley.edu

Cho-Jui Hsieh  
UC Davis  
chohsieh@ucdavis.edu

Richard Vuduc  
Georgia Tech  
richie@cc.gatech.edu

## ABSTRACT

Kernel Ridge Regression (KRR) is a fundamental method in machine learning. Given an  $n$ -by- $d$  data matrix as input, a traditional implementation requires  $\Theta(n^2)$  memory to form an  $n$ -by- $n$  kernel matrix and  $\Theta(n^3)$  flops to compute the final model. These time and storage costs prohibit KRR from scaling up to large datasets. For example, even on a relatively small dataset (a 520k-by-90 input requiring 357 MB), KRR requires 2 TB memory just to store the kernel matrix. The reason is that  $n$  is usually much larger than  $d$  for real-world applications. On the other hand, weak scaling becomes a problem: if we keep  $d$  and  $n/p$  fixed as  $p$  grows ( $p$  is # machines), the memory needed grows as  $\Theta(p)$  per processor and the flops as  $\Theta(p^2)$  per processor. In the perfect weak scaling situation, both the memory needed and the flops grow as  $\Theta(1)$  per processor (i.e. memory and flops are constant). The traditional Distributed KRR implementation (DKRR) only achieved 0.32% weak scaling efficiency from 96 to 1536 processors.

We propose two new methods to address these problems: the Balanced KRR (BKRR) and K-means KRR (KKRR). These methods consider alternative ways to partition the input dataset into  $p$  **different** parts, generating  $p$  **different** models, and then selecting the **best** model among them. Compared to a conventional implementation, KKRR2 (optimized version of KKRR) improves the weak scaling efficiency from 0.32% to 38% and achieves a 591 $\times$  speedup for getting the same accuracy by using the same data and the same hardware (1536 processors). BKRR2 (optimized version of BKRR) achieves a higher accuracy than the current fastest method using less training time for a variety of datasets. For the applications requiring only approximate solutions, BKRR2 improves the weak scaling efficiency to 92% and achieves 3505 $\times$  speedup (theoretical speedup: 4096 $\times$ ).

## CCS CONCEPTS

• **Computing methodologies**  $\rightarrow$  **Massively parallel algorithms**;

## KEYWORDS

Distributed Machine Learning, Scalable Algorithm

## ACM Reference Format:

Yang You, James Demmel, Cho-Jui Hsieh, and Richard Vuduc. 2018. Accurate, Fast and Scalable Kernel Ridge Regression on Parallel and Distributed Systems. In *Proceedings of The 32nd ACM International Conference on Supercomputing (ICS'18)*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3205289.3205290>

ICS'18, June 12–15, 2018, Beijing, China

2018. ACM ISBN 978-1-4503-5783-8/18/06...\$15.00  
<https://doi.org/10.1145/3205289.3205290>

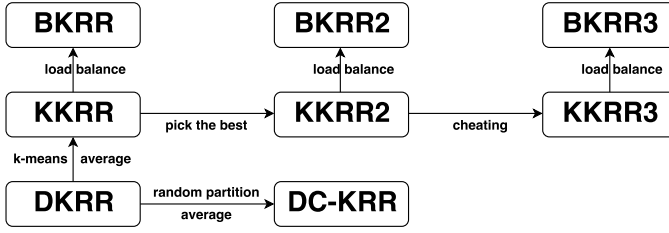
## 1 INTRODUCTION

Learning non-linear relationships between predictor variables and responses is a fundamental problem in machine learning [2], [3]. One state-of-the-art method is Kernel Ridge Regression (KRR) [23], which we target in this paper. It combines ridge regression, a method to address ill-posedness and overfitting in the standard regression setting via L2 regularization, with kernel techniques, which adds flexible support for capturing non-linearity.

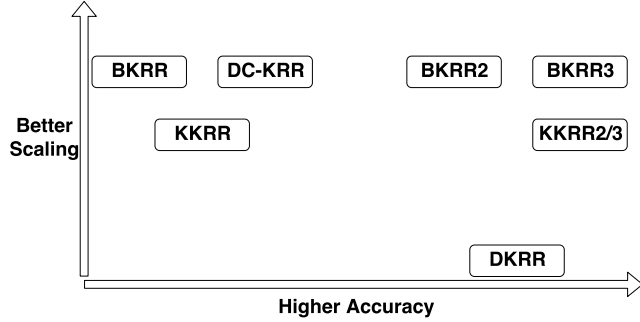
Computationally, the input of KRR is an  $n$ -by- $d$  data matrix with  $n$  training samples and  $d$  features, where typically  $n \gg d$ . At training time, KRR needs to solve a large linear system  $(K + \lambda nI)\alpha = y$  where  $K$  is an  $n$ -by- $n$  matrix,  $\alpha$  and  $y$  are  $n$ -by-1 vectors, and  $\lambda$  is a scalar. Forming and solving this linear system is the major bottleneck of KRR. For example, even on a relatively small dataset—357 megabytes (MB) for a 520,000-by-90 [3]—KRR needs to form a 2 terabyte dense kernel matrix. A standard distributed parallel dense linear solver on a  $p$ -processor system will require  $\Theta(n^3/p)$  arithmetic operations per processor; we refer to this approach hereafter as Distributed KRR (DKRR). In machine learning, where weak scaling is of primary interest, DKRR fares poorly: keeping  $n/p$  fixed, the total storage grows as  $\Theta(p)$  per processor and the total flops as  $\Theta(p^2)$  per processor. In the perfect weak scaling situation, both the memory needed and the flops grow as  $\Theta(1)$  per processor (i.e. memory and flops are constant). In one experiment, the weak scalability of DKRR dropping from 100% to 0.32% as  $p$  increased from 96 to 1536 processors.

Divide-and-Conquer KRR (DC-KRR) [23], addresses the scaling problems of DKRR. Its main idea is to randomly shuffle the rows of the  $n$ -by- $d$  data matrix and then partition it to  $p$  different  $n/p$ -by- $d$  matrices, one per machine, leading to an  $n/p$ -by- $n/p$  kernel matrix on each machine; it builds local KRR models that are then reduced globally to obtain the final model. DC-KRR reduced the memory and computational requirement. However, it can't be used in practice because it sacrifices accuracy. For example, on the Million Song Dataset (a recommendation system application) with 2k test samples, the mean squared error (MSE) decreased only from 88.9 to 81.0 as the number of training samples increased from 8k to 128k. We double the number of processors as we double the number of samples. This is a bad weak scaling problem in terms of accuracy. By contrast, the MSE of DKRR decreased from 91 to 0.002, which is substantially better.

The idea of DC-KRR is to partition the dataset into  $p$  **similar** parts and generate  $p$  **similar** models, and then **average** these  $p$  models to get the final solution. We seek other ways to partition the problem that scale as well as DC-KRR while improving its accuracy. Our idea is to partition the input dataset into  $p$  **different** parts and generate  $p$  **different** models, from which we then select the **best** model among them. Further addressing particular communication



**Figure 1: Optimization flow of our algorithm. DKRR is the baseline. DC-KRR is the existing method. All the others are the new approaches proposed in this paper.**



**Figure 2: Trade-off between accuracy and speed for large-scale weak scaling study. BKRR3 and KKRR3 are the unrealistic approaches. BKRR2 is optimal for scaling and has good accuracy. KKRR2 is optimal for accuracy and has good speed.**

overheads, we obtain two new methods, which we call Balanced KRR (BKRR) and K-means KRR (KKRR). Figure 1 is the summary of our optimization flow. We proposed a series of approaches with details explained later (KKRR3 is an impractical algorithm used later to bound the attainable accuracy). Figure 2 shows the fundamental trade-off between accuracy and scaling for these approaches. Among them, we recommend BKRR2 (optimized version of BKRR) and KKRR2 (optimized version of KKRR) to use in practice. BKRR2 is optimized for scaling and has good accuracy. KKRR2 is optimized for accuracy and has good speed.

When we increase the number of samples from 8k to 128k, KKRR2 (optimized version of KKRR) reduces the MSE from 95 to  $10^{-7}$ , which addresses the poor accuracy of DC-KRR. In addition, KKRR2 is faster than DC-KRR for a variety of datasets. Our KKRR2 method improves weak scaling efficiency over DKRR from 0.32% to 38% and achieves a 591× speedup over DKRR on the same data at the same accuracy and the hardware (1536 processors). For the applications requiring only approximate solutions, BKRR2 improves the weak scaling efficiency to 92% and achieves 3505× speedup with only a slight loss in accuracy.

## 2 BACKGROUND

### 2.1 Linear Regression and Ridge Regression

In machine learning, linear regression is a widely-used method for modeling the relationship between a scalar dependent variable  $y$

(regressand) and multiple explanatory variables (independent variables) denoted by  $x$ , where  $x$  is a  $d$ -dimensional vector. The training data of a linear regression problem consists of  $n$  data points, where each data point (or training sample) is a pair  $(x_i, y_i)$  and  $1 \leq i \leq n$ . Linear regression has two phases: training and prediction. The training phase builds a model from an input set of training samples, while the prediction phase uses the model to predict the unknown regressands  $\hat{y}$  of a new data point  $\hat{x}$ . The training phase is the main limiter to scaling, both with respect to increasing the training set size  $n$  and number of processors  $p$ . In contrast, prediction is embarrassingly parallel and cheap per data point.

**Training Process.** Given  $n$  training data points  $\{(x_i, y_i)\}_{i=1}^n$  where each  $x_i = (x_i^1, \dots, x_i^j, \dots, x_i^d)$  and  $y_i$  is a scalar, we want to build a model relating a given sample  $(x_i)$  to its measured regressand  $(y_i)$ . For convenience, we define  $X$  as an  $n$ -by- $d$  matrix with  $X_{ij} = x_i^j$  and  $y = (y_1, \dots, y_n)$  be an  $n$ -dimensional vector. The goal of regression is to find a  $w$  such that  $y_i \approx w^T \cdot x_i$ . This can be formulated as a least squares problem in Equation (1). To solve some ill-posed problems or prevent overfitting, L2 regularization [13] is used, as formulated in Equation (2), which is called **Ridge Regression**.  $\lambda$  is a positive parameter that controls  $w$ : the larger is  $\lambda$ , the smaller is  $\|w\|_2$ .

$$\argmin_w \|Xw - y\|_2^2 \quad (1)$$

$$\argmin_w \{\|Xw - y\|_2^2 + \lambda \|w\|_2^2\} \quad (2)$$

**Prediction Process.** Given a new sample  $\hat{x}$ , we can use the  $w$  computed in the training process to predict the regressand of  $\hat{x}$  by  $\hat{y} = w^T \cdot \hat{x}$ . If we have  $k$  test samples  $\{\hat{x}_i\}_{i=1}^k$  and their true regressands  $\{\hat{y}_i\}_{i=1}^k$ , the accuracy of the estimated regressand  $\{\hat{y}_i\}_{i=1}^k$  can be evaluated by MSE (Mean Squared Error) defined in Equation (3).

$$MSE = \frac{1}{k} \sum_{i=1}^k (\hat{y}_i - \hat{y}_i)^2 \quad (3)$$

### 2.2 Kernel Method

For many real problems, the underlying model cannot be described by a linear function, so linear ridge regression suffers from poor prediction error. In those cases, a common approach is to map samples to a high dimensional space using a nonlinear mapping, and then learn the model in the high dimensional space. Kernel method [8] is a widely used approach to conduct this learning procedure implicitly by defining the kernel function—the similarity of samples in the high dimensional space. The commonly used kernel functions are shown in Table 1, and we use the most widely used Gaussian kernel in this paper.

**Table 1: Standard Kernel Functions**

Linear	$\Phi(x_i, x_j) = x_i^T x_j$
Polynomial	$\Phi(x_i, x_j) = (ax_i^T x_j + r)^d$
Gaussian	$\Phi(x_i, x_j) = \exp(-\ x_i - x_j\ ^2 / (2\sigma^2))$
Sigmoid	$\Phi(x_i, x_j) = \tanh(ax_i^T x_j + r)$

## 2.3 Kernel Ridge Regression (KRR)

Combining the Kernel method with Ridge Regression yields Kernel Ridge Regression, which is presented in Equations (4) and (5). The  $\|\cdot\|_H$  in Equation (4) is a Hilbert space norm [23]. Given the  $n$ -by- $n$  kernel matrix  $K$ , this problem reduces to a linear system defined in Equation (6).  $K$  is the kernel matrix constructed from training data by  $K_{i,j} = \Phi(x_i, x_j)$ ,  $y$  is the input  $n$ -by-1 regressand vector corresponding to  $X$ , and  $\alpha$  is the  $n$ -by-1 unknown solution vector.

$$\operatorname{argmin} \frac{1}{n} \sum_{i=1}^n \|f_i - y_i\|_2^2 + \lambda \|f\|_H^2 \quad (4)$$

$$f_i = \sum_{j=1}^n \alpha_j \Phi(x_j, x_i) \quad (5)$$

In the Training phase, the algorithm's goal is to get  $\alpha$  by (approximately) solving the linear system in (6). In the Prediction phase, the algorithm uses  $\alpha$  to predict the regressand of any unknown sample  $\hat{x}$  using Equation (7). KRR is specified in Algorithm 1. The algorithm searches for the best  $\sigma$  and  $\lambda$  from parameter sets. Thus, in practice, Algorithm 1 is only a single iteration of KRR because people do not know the best parameters before using the dataset.  $|\Lambda| \times |\Sigma|$  is the number of iterations where  $\Lambda$  and  $\Sigma$  are the parameter sets of  $\lambda$  and  $\sigma$  (Gaussian Kernel) respectively. Thus, the computational cost of KRR method is  $\Theta(|\Lambda||\Sigma|n^3)$ . In a typical case, if  $|\Lambda| = 50$  and  $|\Sigma| = 50$ , then the algorithm needs to finish thousands of iterations. People also use cross-validation technique to select the best parameters, which needs much more time.

$$(K + \lambda n I) \alpha = y \quad (6)$$

$$\hat{y} = \sum_{i=1}^n \alpha_i \Phi(x_i, \hat{x}) \quad (7)$$

---

### Algorithm 1: Kernel Ridge Regression (KRR)

---

#### Input:

$n$  labeled data points  $(x_i, y_i)$  for training;  
 another  $k$  labeled data points  $(\hat{x}_j, \hat{y}_j)$  for testing;  
 both  $x_i$  and  $\hat{x}_j$  are  $d$ -dimensional vectors;  
 $i \in \{1, 2, \dots, n\}, j \in \{1, 2, \dots, k\}$ ;  
 tuned parameters  $\lambda$  and  $\sigma$

#### Output:

Mean Squared Error (MSE) of prediction

```

1 Create a  $n$ -by- $n$  kernel matrix  $K$ 
2 for  $i \in 1 : n$  do
3   for  $j \in 1 : n$  do
4      $K[i][j] \leftarrow \Phi(x_i, x_j)$  based on Table 1
5 Solve linear equation  $(K + \lambda n I) \alpha = y$  for  $\alpha$ 
6 for  $j \in 1 : k$  do
7    $\hat{y}_j \leftarrow \sum_{i=1}^n \alpha_i K(x_i, \hat{x}_j)$ 
8  $MSE \leftarrow \frac{1}{k} \sum_{j=1}^k (\hat{y}_j - \hat{y}_j)^2$ 
```

---

## 2.4 K-means clustering

Here we review K-means clustering algorithm, which will be used in our algorithm. The objective of K-means clustering is to partition a dataset  $TD$  into  $k \in \mathbb{Z}^+$  subsets  $(TD_1, TD_2, \dots, TD_k)$ , using a notion of proximity based on Euclidean distance [7]. The value of  $k$  is chosen by the user. Each subset has a center  $(CT_1, CT_2, \dots, CT_k)$ ,

each of which is a  $d$ -dimensional vector. A sample  $x$  belongs to  $TD_i$  if  $CT_i$  is its closest center. K-means is shown in Algorithm 2.

---

### Algorithm 2: Plain K-means Clustering

---

```

1 Input the training samples  $x_i, i \in \{1, 2, \dots, n\}$ 
2 Initialize data centers  $CT_1, CT_2, \dots, CT_k$  randomly
3  $\delta \leftarrow 0$ 
4 For every  $i = 1, \dots, n$ 
5    $c^i \leftarrow \operatorname{argmin}_j \|x_i - CT_j\|$ 
6   If  $c^i$  has been changed,  $\delta \leftarrow \delta + 1$ 
7 End For
8 For every  $j = 1, \dots, k$ 
9    $CT_j \leftarrow \frac{\sum_{i=1}^n 1\{c^i=j\} x_i}{\sum_{i=1}^n 1\{c^i=j\}}$ 
10 End For
11 If  $\delta/n > \text{threshold}$ , then go to Step 3
```

---

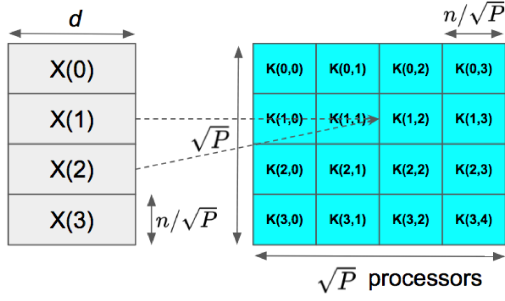
## 3 EXISTING METHODS

### 3.1 Distributed KRR (DKRR)

The bottleneck of KRR is solving the  $n$ -by- $n$  linear system (6), which is generated by a much smaller  $n$ -by- $d$  input matrix with  $n \gg d$ . As stated before, this makes weak-scaling problematic, because memory-per-machine grows like  $\Theta(p)$ , and flops-per-processor grows like  $\Theta(p^2)$ . In the perfect weak scaling situation, both the memory needed and the flops grow as  $\Theta(1)$  per processor (i.e. memory and flops are constant). Our experiments show the weak scaling efficiency of DKRR decreases from 100% to 0.3% as we increase the number of processors from 96 to 1536. Since the  $n$ -by- $n$  matrix  $K$  cannot be created on a single node, we create it in a distributed way on a  $\sqrt{p}$ -by- $\sqrt{p}$  machine grid (Fig. 3). We first divide the sample matrix into  $\sqrt{p}$  equal parts by rows. To generate  $\frac{1}{p}$  of the kernel matrix, each machine will need two of these  $\sqrt{p}$  parts of the sample matrix. For example, in Fig. 3, to generate the  $K(1,2)$  block, machine 6 needs the second and the third parts of the blocked sample matrix. Thus, we reduce the storage and computation for kernel creation from  $\Theta(n^2)$  to  $\Theta(n^2/p)$  per machine. Then we use distributed linear solver in ScaLAPACK [5] to solve for  $\alpha$ .

### 3.2 Divide-and-Conquer KRR (DC-KRR)

DC-KRR [23] showed that using divide-and-conquer can reduce the memory and computational requirement. We first shuffle the original data matrix  $M = [X, y]$  by rows. Then we distribute  $M$  to all the nodes evenly by rows (lines 1-5 of Algorithm 3). On each node, we construct a much smaller matrix ( $\Theta(n^2/p^2)$ ) than the original kernel matrix ( $\Theta(n^2)$ ) (lines 6-11 of Algorithm 3). After getting the local kernel matrix  $K$ , we use it to solve a linear equation  $(K + \lambda n I) \alpha = y$  on each node where  $y$  is the input labels and  $\alpha$  is the solution. After getting  $\alpha$ , the training step is complete. Then we use the local  $\alpha$  to make predictions for each unknown data point and do a global average (reduction) to output the final predicted label (lines 13-15 of Algorithm 3). If we get a set of better hyper-parameters, we record them by overwriting the old versions (lines 16-19 of Algorithm 3).



**Figure 3: Implementation of Distributed KRR.** We divide the input sample matrix into  $\sqrt{p}$  parts by rows, and each machine gets two of these  $\sqrt{p}$  parts. Then each machine generates  $1/p$  part of the kernel matrix.

---

**Algorithm 3: Divide-and-Conquer KRR (DC-KRR)**


---

**Input:**

$n$  labeled data points  $(x_i, y_i)$  for training;  
 $k$  labeled data points  $(\hat{x}_j, \hat{y}_j)$  for testing;  
 $x_i$  and  $\hat{x}_j$  are  $d$ -dimensional vectors;  
 $i \in \{1, 2, \dots, n\}, j \in \{1, 2, \dots, k\}$ ;  
 $\overline{MSE} \leftarrow \infty$  (Initial Mean Squared Error);

**Output:**

Mean Squared Error ( $\overline{MSE}$ ) of prediction;  
 best parameters  $\hat{\lambda}, \hat{\sigma}$ ;

```

1 if rank = 0 then
2   Store data points  $(x_i, y_i)$   $i \in \{1, \dots, n\}$  as  $[X, y]$ 
3    $M = [X, y]$  is a  $n$ -by- $(d+1)$  matrix
4   Shuffle  $M$  by rows
5   Scatter  $M$  to all the nodes evenly by rows
6 Create a  $m$ -by- $m$  kernel matrix  $K$ ,  $m = n/p$ 
  for  $i \in 1 : m$  do
7    $x_i, y_i = M[i][1:d+1]$ 
8 for  $\lambda \in \Lambda$  and  $\sigma \in \Sigma$  do
9   for  $i \in 1 : m$  do
10    for  $j \in 1 : m$  do
11       $K[i][j] = \Phi(x_i, x_j)$  based on Table 1
12   Solve linear equation  $(K + \lambda m I)\alpha = y$  for  $\alpha$ 
13   for  $j \in 1 : k$  do
14      $\hat{y}_j = \sum_{i=1}^m \alpha_i K(x_i, \hat{x}_j)$ 
15   Global reduce:  $\tilde{y} = \sum \hat{y}/p$ 
16   if rank = 0 then
17      $MSE = \frac{1}{k} \sum_{j=1}^k (\tilde{y}_j - \hat{y}_j)^2$ 
18     if  $MSE < \overline{MSE}$  then
19        $\overline{MSE} \leftarrow MSE, \hat{\lambda} \leftarrow \lambda, \hat{\sigma} \leftarrow \sigma$ 

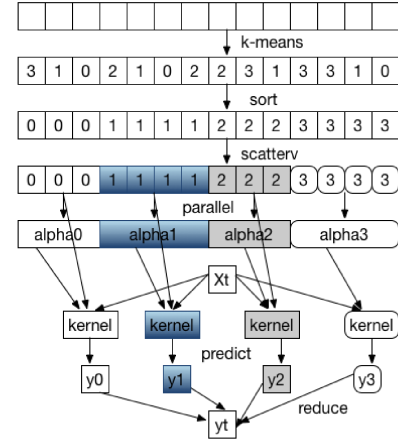
```

---

## 4 ACCURATE, FAST, AND SCALABLE KRR

### 4.1 Kmeans KRR (KKRR)

DC-KRR performs better than state-of-the-art methods [23]. However, based on our observation, DC-KRR still needs to be improved. DC-KRR has a poor weak scaling in terms of accuracy, which is the bottleneck for distributed machine learning workloads. We observe

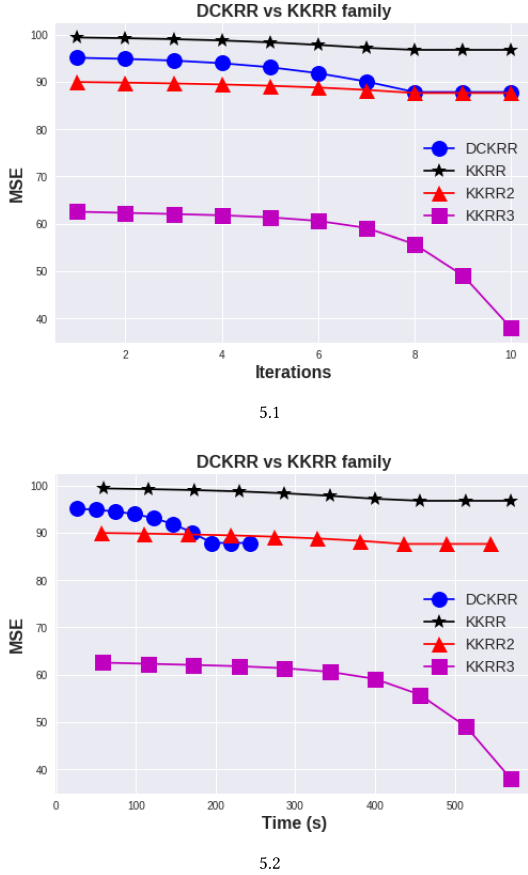


**Figure 4: Implementation of Kmeans KRR (KKRR).** Both k-means and sort can be parallelized. We use the standard MPI implementation for scatter operation.

that the poor accuracy of DC-KRR is mainly due to the random partitioning of training samples. Thus, our objective is to design a better partitioning method to achieve a higher accuracy and maintain a high scaling efficiency. Our algorithm is accurate, fast, and scalable on distributed systems.

The analysis in this section is based on the Gaussian kernel because it is the most widely used case [23]. Other kernels can work in the same way with different distance metrics. For any two training samples, their kernel value ( $\exp\{-\|x_i - x_j\|^2/(2\sigma^2)\}$ ) is close to zero when their Euclidean distance ( $\|x_i - x_j\|^2$ ) is large. Therefore, for a given sample  $\hat{x}$ , only the training points close to  $\hat{x}$  in Euclidean distance can have an effect on the result (Equation (7)) of the prediction process. Based on this idea, we partition the training dataset into  $p$  subsets ( $TD_1, TD_2, \dots, TD_p$ ). We use K-means to partition the dataset since K-means maximizes Euclidean distance between any two clusters. The samples with short Euclidean distance will be clustered into one group. After K-means, each subset will have its data center ( $CT_1, CT_2, \dots, CT_p$ ). Then we launch  $p$  independent KRRs ( $KRR_1, KRR_2, \dots, KRR_p$ ) to process these  $p$  subsets.

Given a test sample  $\hat{x}$ , we predict of only using one model to predict  $\hat{x}$ , we make all the nodes have a copy of  $\hat{x}$ . This additional cost is trivial for two reasons: (1) the test dataset is much smaller than the training dataset, and the training dataset is much smaller than the kernel matrix, which is the major memory overhead. (2) This broadcast is finished at initial step and there is no need to do it at every iteration. Each node  $i$  makes a prediction  $\hat{y}_i$  for  $\hat{x}$  by using its local model. Then we get the final prediction  $\tilde{y}$  by a global reduction (average) over all nodes ( $\tilde{y} = \sum_{i=1}^p \hat{y}_i/p$ ). Figure 4 is the framework of KKRR. KKRR is highly parallel because all the subproblems are independent of each other. However, the performance of KKRR is not good as we expect. From Figure 5.1, we observe that DC-KRR achieves better accuracy than KKRR. In addition, KKRR is slower than DC-KRR. In the following sections, we will make the algorithm more accurate and scalable, which is better than DC-KRR.



**Figure 5: Comparison between DC-KRR and KKRR family, using same parameter set on 96 NERSC Edison processors. The test data set MSD is described in Section 5.1. KKRR2 is an accurate but slow algorithm. KKRR3 is an optimal but unrealistic method for comparison purposes.**

## 4.2 KKRR2

The low accuracy of KKRR is mainly due to its conquer step. For DC-KRR, because the divide step is a random and even partition, the sub-datasets and local models are similar to each other, and thus averaging works pretty well. For KKRR, the clustering method divides the original dataset into different sub-datasets which are far away from each other in Euclidean distance. The local models generated by sub-datasets are also totally different from each other. Thus, using their average will get worse accuracy because some models are unrelated to the test sample  $\hat{x}$ . For example, if the data center of  $i$ -th partition ( $CT_i$ ) is far away from  $\hat{x}$  in Euclidean distance, the  $i$ -th model should not be used to make prediction for  $\hat{x}$ . Since we divide the original dataset based on Euclidean distance, the similar procedure should be used in the prediction phase. Thus, we design the following algorithm.

After the training process, each sub-KRR will generate its own model file ( $MF_1, MF_2, \dots, MF_p$ ). We can use these models independently for prediction. For a given unknown sample  $\hat{x}$ , if its closest

data center (in Euclidean distance) is  $CT_i$ , we only use  $MF_i$  to make a prediction for  $\hat{x}$ . We call this version KKRR2. From Figures 5.1 and 5.2 we observe that KKRR2 is more accurate than DCKRR. However, KKRR2 is slower than DCKRR. For example, to get the same accuracy (MSE=88), KKRR2 is 2.2× (436s vs 195s) slower than DCKRR. Thus we need to focus on speed and scalability.

## 4.3 Balanced KRR (BKRR)

Based on the profiling results in Figure 6, we observe that the major source of KKRR's poor efficiency is load imbalance. The reason is that the partitioning by K-means is irregular and imbalanced. For example, processor 2 in Figure 6 has to handle  $n = 35,137$  samples while processor 3 only needs to process  $n = 7,349$  samples. Since the memory requirement grows like  $\Theta(n^2)$  and the number of flops grows like  $\Theta(n^3)$ , processor 3 runs 51× faster than processor 1 (Figure 6). On the other hand, partitioning by K-means is data-dependent and the sizes of clusters cannot be controlled. This makes it unreliable to be used in practice, and thus we need to replace K-means with a load-balanced partitioning algorithm. To this end, we design a K-balance clustering algorithm and use it to build Balanced Kernel Ridge Regression (BKRR).

---

### Algorithm 4: K-balance Clustering

---

**Input:**

$CT[i]$  is the center of  $i$ -th cluster;  
 $CS[i]$  is the size of  $i$ -th cluster;  
 $SA[i]$  is the  $i$ -th sample;  
 $n$  is the number of samples;  
 $P$  is the number of clusters (#machines);

**Output:**

$MB[i]$  is the closest center to  $i$ -th sample;  
 $CT[i]$  is the center of  $i$ -th cluster;

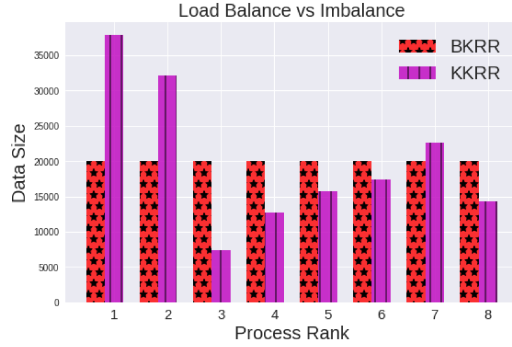
```

1 Run K-means to get  $CT[1], CT[2], \dots, CT[p]$ 
2  $balanced = n/p$ 
3 for  $i \in 1 : n$  do
4    $mindis = \inf$ 
5    $minind = 0$ 
6   for  $j \in 1 : p$  do
7      $dist = \text{EuclidDistance}(SA[i], CT[j])$ 
8     if  $dist < mindis$  and  $CS[j] < balanced$  then
9        $mindis = dist$ 
10       $minind = j$ 
11    $CS[minind]++$ 
12    $MB[i] = minind$ 
13 for  $i \in 1 : p$  do
14    $CT[i] = 0$ 
15 for  $i \in 1 : n$  do
16    $j = MB[i]$ 
17    $CT[j] += SA[i]$ 
18 for  $i \in 1 : p$  do
19    $CT[i] = CT[i] / CS[i]$ 

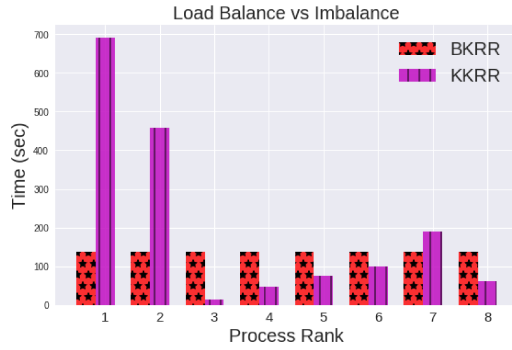
```

---

In our design, a machine corresponds to a clustering center. If we have  $p$  machines, then we partition the training dataset into  $p$  parts. As mentioned above, the objective of K-balance partitioning algorithm is to make the number of samples on each node close



6.1 Load Balance for Data Size



6.2 Load Balance for Time

**Figure 6:** This experiment is conducted on NERSC Cori Supercomputer [19]. We use 8 nodes for load balance test. The test dataset is MSD dataset and we use 16000 samples. We observe that BKRR has roughly 2000 samples on each node, which leads to a perfect load balance. For KKRR, because different nodes have different number of samples, the fastest node is 51× faster than the slowest node, which leads to a huge resource waste.

to  $n/p$ . If a data center has  $n/p$  samples, then we say it is balanced. The basic idea of this algorithm is to find the closest center for each sample, and if a given data center has been balanced, no additional sample will be added to this center. The detailed K-balance clustering method is in Algorithm 4. Line 1 of Algorithm 4 is an important step because K-balance needs to first run K-means algorithm to get the data centers. This makes K-balance have a similar clustering pattern as K-means. Lines 3-12 find the center for each sample. Lines 6-10 find the best under-load center for the  $i$ -th sample. Lines 15-19 recompute the data center by averaging all the samples in a certain center. Recomputing the centers by averaging is optional because it will not necessarily make the results better. From Figure 6 we observe that K-balance partitions the dataset in a balanced way. After K-balance clustering, all the nodes have the same number of samples, so in the training phase all the nodes roughly have the same training time and memory requirement.

After replacing K-means with K-balance, KKRR becomes BKRR, and KKRR2 becomes BKRR2. Algorithm 5 is a framework of BKRR2.

---

**Algorithm 5:** Balanced KRR2 (BKRR2)

---

**Input:**

$n$  labeled data points  $(x_i, y_i)$  for training;  
 $k$  labeled data points  $(\hat{x}_j, \hat{y}_j)$  for testing;  
both  $x_i$  and  $\hat{x}_j$  are  $d$  dimensional vectors;  
 $i \in \{1, 2, \dots, n\}, j \in \{1, 2, \dots, k\}$ ;  
 $MSE \leftarrow \infty$  (Initial Mean Squared Error);

**Output:**

Mean Squared Error ( $\widehat{MSE}$ ) of prediction  
best parameters  $\hat{\lambda}, \hat{\sigma}$

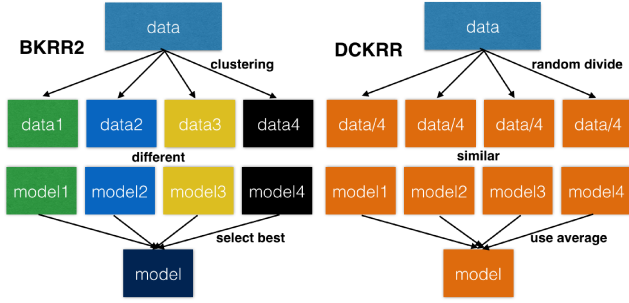
```

1  $t \leftarrow$  rank of a machine,  $t \in \{1, \dots, p\}$ 
2 Do K-balance clustering (Algorithm 4)
3  $M = [X, y]$  is a  $(n/p)$ -by- $(d+1)$  matrix
4 Store data points  $(x_i, y_i)$   $i \in \{1, \dots, n/p\}$  as  $[X, y]$ 
5 Create a  $m$ -by- $m$  kernel matrix  $K$ ,  $m = n/p$ 
   for  $i \in 1 : m$  do
6    $x_i, y_i = M[i][1:d+1]$ 
7   Machine  $t: (x_i^t, y_i^t) = (x_i, y_i)$ ,  $t \in \{1, \dots, p\}$ 
8 for  $\lambda \in \Lambda$  and  $\sigma \in \Sigma$  do
9   for  $i \in 1 : m$  do
10    for  $j \in 1 : m$  do
11       $K[i][j] = \Phi(x_i, x_j)$  based on Table 1
12    Solve linear equation  $(K + \lambda m I)\alpha = y$  for  $\alpha$ 
13    for  $t \in 1 : p$  do
14       $err_t \leftarrow 0$ 
15    for  $j \in 1 : k$  do
16      if  $t = MyCluster(\hat{x}_j)$  then
17         $\hat{y}_j \leftarrow \sum_{i=1}^m \alpha_i^t K(x_i^t, \hat{x}_j)$ 
18         $err_t \leftarrow err_t + ||\hat{y}_j - \hat{y}_j||^2$ 
19    Reduce:  $MSE = (\sum_{t=1}^p err_t)/k$ 
20    if  $t = 0$  then
21      if  $MSE < \widehat{MSE}$  then
22         $\widehat{MSE} \leftarrow MSE, \hat{\lambda} \leftarrow \lambda, \hat{\sigma} \leftarrow \sigma$ 
```

---

As we mentioned in the sections of Abstract and Introduction, KKRR2 is the optimized version of KKRR and BKRR2 is the optimized version of BKRR. Lines 1-7 perform the partition. Lines 9-22 perform one iteration of the training. Lines 9-11 construct the kernel matrix. Line 12 solves the linear equation on each machine. Lines 13-18 perform the prediction. Only the best model does the prediction for a certain test sample. Lines 19-22 evaluate the accuracy of the system. There is no communication cost of BKRR2 during training except for the initial data partition and the final model selection. The additional cost of BKRR2 comes from two parts: (1) the K-balance algorithm and partition, and (2) the prediction.

The overhead of K-balance is tiny compared with the training part. K-balance first does K-means. The cost of K-means is  $\Theta(In)$  where  $I$  is the number of K-means iterations, which is usually less 100. The cost of lines 3-12 in Algorithm 4 is  $\Theta(pn)$  where  $p$  is the number of partitions and also the number of machines.  $\Theta(In + pn)$  is tiny compared with the training part, which is  $\Theta(|\Sigma||\Lambda|n^3/p^3)$ . For example, if we use BKRR2 to process the 32k MSD training samples on 96 NERSC Edison processors, the single-iteration training time



**Figure 7: Difference between BKRR2 and DCKRR. BKRR2: partition the dataset into  $p$  different parts, generate  $p$  different models, and select the best model. DCKRR: partition the dataset into  $p$  similar parts, generate  $p$  similar models, and use the average of all the models.**

is 20 times larger than the K-balance and partition time. Since the computational overhead of K-balance is low compared to KRR training ( $n = \Theta(p^4)$  in practice), we can just use one node to finish the K-balance algorithm. Although it is not necessary, we have an approximate parallel algorithm for K-balance. We conduct parallel K-means clustering and distribute the centers to all the nodes. Then we partition the samples to all the processors and set *balanced* as  $n/p^2$ . Since  $n = \Theta(p^4)$  in practice, this parallel implementation roughly gets the same results with the single-node implementation. The overhead of this approximate parallel algorithm is  $\Theta(\ln/p + n)$ .

For the prediction part, instead of conducting  $k$  small communications, we make each machine first compute its own error (line 18 of Algorithm 5) and then only conduct one global communication (line 19 of Algorithm 5). The reason is detailed below. The runtime on a distributed system can be estimated by the sum of the time for computation and the time for communication:  $f \times \gamma + n_m \times (\alpha + n_b \times \beta)$ , where  $f$  is the number of floating point operations,  $\gamma$  is the time per floating point operation,  $\alpha$  is latency,  $1/\beta$  is the bandwidth,  $n_m$  is the number of messages, and  $n_b$  is the number of bytes per message. In practice,  $\alpha \gg \beta \gg \gamma$  (e.g.  $\alpha = 7.2 \times 10^{-6}$ s,  $\beta = 0.9 \times 10^{-9}$ s for Intel NetEffect NE020,  $\gamma = 2 \times 10^{-11}$ s for Intel 5300-series Clovertown). Therefore, one big message is much cheaper than  $k$  small messages because  $k \times (\alpha + n_b \times \beta) \gg (\alpha + k \times n_b \times \beta)$ . This optimization reduces the latency overhead. Figures 8 and 9 show that BKRR2 achieves lower error rate than DC-KRR in a shorter time for a variety of datasets. Figure 7 shows the difference between BKRR2 and DC-KRR.

#### 4.4 BKRR3 and KKRR3: Error Lower Bound and The Unrealistic Approach

The KKRR and BKRR families share the same idea of partitioning the data into  $p$  parts and making  $p$  different models, but they are different in that the KKRR family is optimized for accuracy while the BKRR family is optimized for performance. We want to know the gap between our method and the best theoretical method. Let us refer to the theoretical KKRR method as KKRR3 and the theoretical BKRR method as BKRR3.

BKRR3 is similar to BKRR2 in terms of communication and computation pattern. Like BKRR and BKRR2, after the training process, each sub-BKRR will generate its own model file ( $MF_1, MF_2, \dots, MF_p$ ). We can use these model files independently for prediction. For a given test sample  $\hat{x}$  ( $\hat{y}$  is its true regressand), we make all the nodes have a copy of  $\hat{x}$  (like KKRR). Each model will make a prediction for  $\hat{x}$ . We get  $\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_p$  from  $p$  models respectively. We select  $MF_i$  for prediction where  $i = \arg\min_j ||\tilde{y}_j - \hat{y}||^2$ . This means we **inspect** the true regressand to make sure we select the best model for each test sample. When changing K-balance to K-means, BKRR3 becomes KKRR3, which is much more accurate than DCKRR (Figures 5.1 and 5.2). The MSE of BKRR3 is the lower bound of the MSE of BKRR2 because BKRR3 can always use the best model for each test sample. In fact, BKRR3 is even much more accurate than the original method (DKRR) for all the testings in our experiments. The framework of BKRR3 is in Algorithm 6. Figures 8 and 9 show the results. BKRR3 is always the best approach in these comparisons. BKRR3 achieves much higher accuracy than DCKRR and BKRR2.

---

#### Algorithm 6: BKRR3

---

##### Input:

$n$  labeled data points  $(x_i, y_i)$  for training;  
 another  $k$  labeled data points  $(\hat{x}_j, \hat{y}_j)$  for testing;  
 both  $x_i$  and  $\hat{x}_j$  are  $d$  dimensional vectors;  
 $i \in \{1, 2, \dots, n\}, j \in \{1, 2, \dots, k\}$ ;  
 $t$  is the rank of a machine,  $t \in \{1, \dots, p\}$ ;

##### Output:

Mean Squared Error ( $\widehat{MSE}$ ) of prediction  
 best parameters  $\hat{\lambda}, \hat{\sigma}$

```

1 Do K-balance clustering
2  $M = [X, y]$  is a  $(n/p)$ -by- $(d+1)$  matrix
3 Store data points  $(x_i, y_i) \ i \in \{1, \dots, n/p\}$  as  $[X, y]$ 
4 Create a  $m$ -by- $m$  kernel matrix  $K, m = n/p$ 
   for  $i \in 1 : m$  do
5    $x_i, y_i = M[i][1:d+1]$ 
6   Machine  $t: (x_i^t, y_i^t) = (x_i, y_i), t \in \{1, \dots, p\}$ 
7 for  $\lambda \in \Lambda$  and  $\sigma \in \Sigma$  do
8   for  $i \in 1 : m$  do
9     for  $j \in 1 : m$  do
10       $K[i][j] = \Phi(x_i, x_j)$  based on Table 1
11   Solve linear equation  $(K + \lambda m I)\alpha = y$  for  $\alpha$ 
12   for  $j \in 1 : k$  do
13     Machine  $t: \tilde{y}_j^t = \sum_{i=1}^m \alpha_i K(x_i^t, \hat{x}_j)$ 
14     Global reduce:  $id = \arg\min_t ||\tilde{y}_j^t - \hat{y}_j||^2$ 
15     Send/Receive:  $\tilde{y}_j = \tilde{y}_j^{id}$ 
16   if  $rank = 0$  then
17      $MSE = \frac{1}{k} \sum_{j=1}^k (\tilde{y}_j - \hat{y}_j)^2$ 
18     if  $MSE < \widehat{MSE}$  then
19        $\widehat{MSE} \leftarrow MSE, \hat{\lambda} \leftarrow \lambda, \hat{\sigma} \leftarrow \sigma$ 

```

---



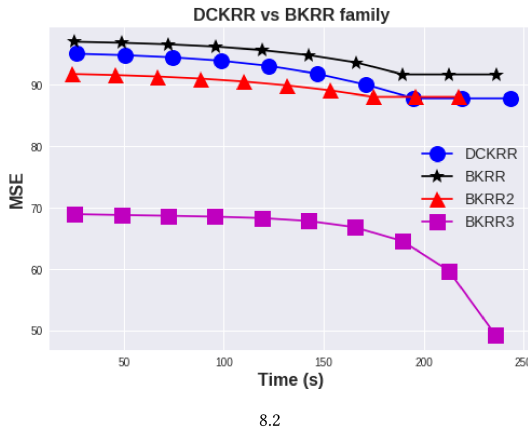
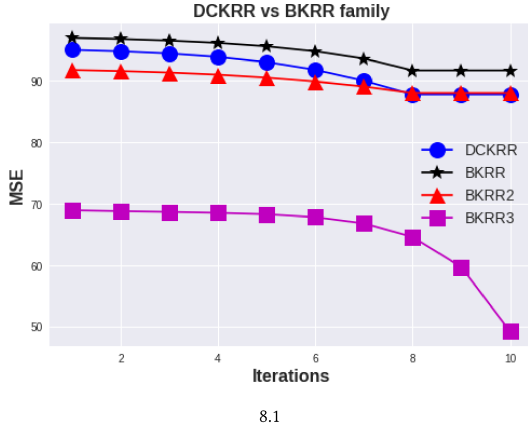


Figure 8: These figures do a point-do-point comparison between DC-KRR and BKRR family using test data set MSD (described in Section 5.1). They use the same parameter set and conduct the same number of iterations on 96 NERSC Edison processors. DCKRR is more accurate than BKRR. BKRR2 is faster than DCKRR for getting the same accuracy. BKRR3 is an optimal but unrealistic implementation for comparison purposes.

## 5 IMPLEMENTATION AND ANALYSIS

### 5.1 Real-World Dataset

To give a fair comparison with DC-KRR, we use the Million Song Dataset (MSD) [3] as our major dataset in our experiments because MSD was used in the paper of DC-KRR. It is a freely-available collection of audio features and metadata for a million contemporary popular music tracks. The dataset contains  $n = 515,345$  samples. Each sample is a song (track) released between 1922 and 2011, and the song is represented as a vector of timbre information computed from the song. Each sample consists of a pair of  $(x_i, y_i)$  where  $x_i$  is a  $d$ -dimensional ( $d = 90$ ) vector and  $y_i \in [1922, 2011]$  is the year that the song was released. The Million Song Dataset Challenge aims at being the best possible offline evaluation of a music recommendation system. It is a large-scale, personalized music recommendation

challenge, where the goal is to predict the songs that a user will listen to, given both the user's listening history and full information (including meta-data and content analysis) for all songs. To justify the efficiency of our approach, we use another three real-world datasets. The information of these four datasets are summarized in Table 2. All these datasets were downloaded from [9].

Table 2: The test datasets

name	MSD	cadata	MG	space-ga
# Train	463,715	18,432	1,024	2,560
# Test	51,630	2,208	361	547
Dimension	90	8	6	6
Application	Music	Housing	Dynamics	Politics

Table 3: Weak Scaling in time. We use 96 processors and 8000 MSD samples as the baseline. Constant time means perfect scaling. BKRR2 has very good scaling efficiency. DKRR's scaling efficiency is poor.

Method	Processors	96	192	384	768	1536
BKRR2	IterTime (s)	1.06	1.06	1.06	1.08	1.15
KKRR2	IterTime (s)	2.60	3.66	4.05	4.23	6.85
DKRR	IterTime (s)	13	75.1	234	1273	4048
BKRR2	Efficiency	1.0	1.0	1.0	0.99	0.92
KKRR2	Efficiency	1.0	0.71	0.64	0.62	0.38
DKRR	Efficiency	1.0	0.17	0.06	0.01	0.003

Table 4: Weak Scaling in Accuracy on MSD Dataset. Lower is better. DCKRR is a bad algorithm.

Samples	DCKRR	BKRR2	DKRR	KKRR2	BKRR3
8k	88.9	93.1	90.9	95.0	40.2
32k	85.5	87.7	85.0	87.5	6.6
128k	81.0	14.7	0.002	$10^{-7}$	$10^{-7}$

### 5.2 Fair Comparison

Let us use  $p$  as the number of partitions or nodes,  $\rho$  as the number of processors. Each node has 24 processors. When we use  $\rho=1536$  processors, we actually divide the dataset into  $p=64$  parts. Each machine generates a local model for BKRR2. To give a fair comparison, we make sure all the comparisons were tuned based on the same parameter set. Different methods may select different best-parameters from the same parameter set to achieve its lowest MSE. From Figures 8, 9 and 11, we clearly observe BKRR2 is faster than DC-KRR and also achieves lower prediction error on all the datasets. In other words, BKRR2 and DC-KRR may use different parameters to achieve their lowest MSEs. The lowest MSE of BKRR2 is lower than the lowest MSE of DC-KRR. On the other hand, BKRR2 is slightly faster than both DC-KRR and BKRR3 (for both single iteration time and overall time). The reason is that each machine only needs to process  $1/p$  of the test samples for prediction. For DC-KRR and BKRR3, each machine needs to process all the test



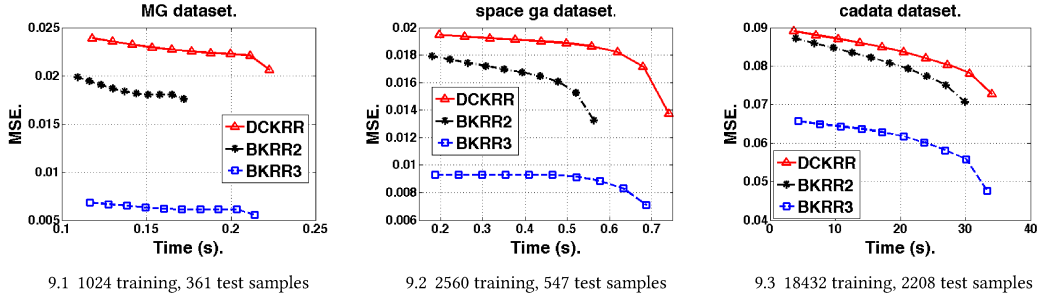


Figure 9: These figures do a point-to-point comparison between BKRR2, BKRR3 and DC-KRR. They use the same parameter set and conduct the same number of iterations on 96 NERSC Edison processors. BKRR2 is faster than DCKRR for getting the same accuracy. BKRR3 is an optimal but unrealistic implementation for comparison purposes.

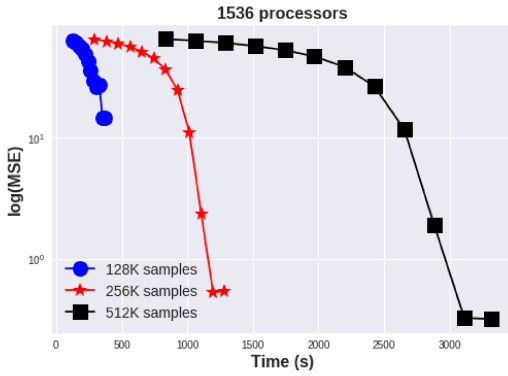


Figure 10: BKRR2 results based on MSD dataset. Using increasing number of samples on a fixed number of processors, we can get a better model but also observe the time increase in the speed of  $\Theta(n^3)$ .

samples for prediction. BKRR2 achieves 1.2 $\times$  speedup over DC-KRR on average and has a much lower error rate for 128k-sample MSD dataset (14.7 vs 81.0).

Because of DKRR's poor weak scaling, BKRR2 runs much faster than DKRR for 1536 processors and 128k training samples. The single iteration time of BKRR2,  $t_b$ , is 1.15 sec while the single iteration time of DKRR,  $t_d$ , is 4048 sec. Here, single iteration means picking a pair of parameters and solving the linear equation once (e.g. lines 9-22 of Algorithm 5 are one iteration). The algorithm can get a better pair of parameters after each iteration. All the algorithms in this paper run the same number of iterations because they use the same parameter tuning space. However, it is unfair to say BKRR2 achieves 3505 $\times$  speedup over DKRR because, for the same 2k test dataset, the lowest MSE of BKRR2 is 14.7 while that of DKRR is 0.002. This means the BKRR2 model with 128K samples model ( $bm_{128}$ ) is worse than DKRR model ( $dm_{128}$ ) for accuracy. To give a fair comparison, we increase the training samples of BKRR2 to 256k so that its lowest MSE can generate a better model ( $bm_{256}$ ). By using  $bm_{256}$ , the lowest MSE of BKRR2 is 0.53 (Figure 10). We run it on the same 1536 processors and observe  $t_b$  becomes 5.6 sec. In this way, we can say that BKRR2 achieves 723 $\times$  speedup

over DKRR for achieving roughly the same accuracy by using the same hardware and the same test dataset. It is worth noting that the biggest dataset that DKRR can handle on 1536 processors is 128k samples, otherwise it will collapse. Therefore,  $dm_{128}$  is the best model DKRR can get on 1536 processors. However, BKRR2 can use a bigger dataset to get a better model than  $bm_{256}$  on 1536 processors (e.g. 512k model in Figure 10).

The theoretical speedup of  $bm_{128}$  over  $dm_{128}$  is the ratio of  $\Theta(n^3/p)$  and  $\Theta((n/p)^3)$ , which is 4096 $\times$  for  $p = 64$  and  $\rho = 1536$ . We achieve 3505 $\times$  speedup. The theoretical speedup of  $bm_{256}$  over  $dm_{128}$  is the ratio of  $\Theta(n^3/p)$  and  $\Theta((2n/p)^3)$ , which is 512 $\times$  for  $p = 64$  and  $\rho = 1536$ . We achieve 723 $\times$  speedup. The difference between theoretical speedup and practical speedup comes from systems and low-level libraries (e.g. the implementation of LAPACK and ScaLAPACK). In this comparison, BKRR2's better scalability allows it to use more data than DKRR, which cannot run on an input of the same size. We also want to compare KKRR2 to DKRR for using the same amount of data. Let us refer to the model generated by KKRR2 using 128k samples as  $km_{128}$  and the single iteration time as  $t_k$ .  $t_k$  is 6.9 sec in our experiment. The MSE of  $km_{128}$  is  $10^{-7}$ , which is even lower than the MSE of  $dm_{128}$  (0.002). Thus, KKRR2 achieves 591 $\times$  speedup over DKRR for the same accuracy by using the same data and hardware (Table 3 and 4).

### 5.3 Weak-Scaling Issue

As we mentioned in a previous section, weak scaling means we fix the data size per node and increase the number of nodes. We use from 96 to 1536 processors (4 to 64 partitions) for scaling tests on the NERSC Edison supercomputer. The test dataset is the MSD dataset. We set the 96-processor case as the baseline and assume it has a 100% weak-scaling. We double the number of samples as we double the number of processors. Figure 11 and Table 3 show the time weak-scaling and accuracy weak-scaling of BKRR. In terms of time weak-scaling, BKRR2 achieves 92% efficiency on 1536 processors while DKRR only achieves 0.32%. We then compare the weak scaling in terms of test accuracy: the MSE of DC-KRR for 2k test samples only decreases from 88.9 to 81.0, while the MSE of BKRR2 decreases from 93.1 to 14.7. KKRR2 reduced the MSE from 95 to  $10^{-7}$ . In conclusion, we observe that DKRR has a bad time scaling and DC-KRR has a poor accuracy weak scaling. Our

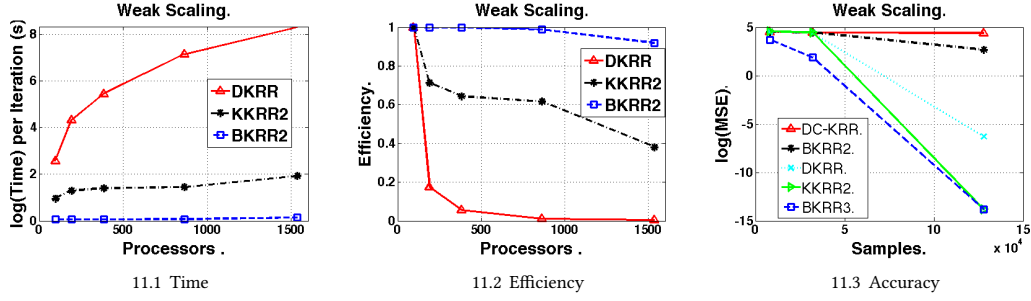


Figure 11: Results based on MSD dataset. We use 96 processors (i.e. 4 nodes) and 8k samples as the baseline. The lowest MSE of DKRR is 0.001848 on 1536 processors. The lowest MSE of BKRR3 is  $10^{-7}$ . The weak scaling efficiency of DKRR at 1536 processors is 0.32%. The weak scaling efficiency of KKRR2 and BKRR2 at 1536 processors is 38% and 92%, respectively. The MSE of DC-KRR for 2k test samples only decreases from 88.9 to 81.0, which is a bad weak scaling accuracy. The MSE of BKRR2 decreases from 93.1 to 14.7. The MSE of KKRR2 decreases from 95.0 to  $10^{-7}$ . The data is in Tables 3 and 4. We conclude that the proposed methods outperform the existing methods.

proposed method, BKRR2, has good weak scaling behavior both in terms of time and accuracy. The weak scaling of BKRR3 can be shown in Table 5.

Table 5: Weak scaling results of BKRR3. We use 96 processors and 8k samples as the baseline. We double the number of samples as we double the number of processors.

processors	96	192	384	768	1536
Time (s)	1311	1313	1328	1345	1471
Efficiency	100%	99.8%	98.7%	97.5%	89.1%
MSE	40.2	16.5	6.62	1.89	$10^{-7}$

## 5.4 Method Selection

KKRR2 achieves better accuracy than the state-of-the-art method. If the users need the most accurate method, KKRR2 is the best choice. KKRR2 runs slower than DC-KRR but much faster than DKRR. BKRR2 achieves better accuracy and also runs faster than the DCKRR. BKRR2 is much faster than DKRR with the same accuracy. The reason why BKRR2 is more efficient than DKRR is not that it assumes more samples, but rather, because it has much better weak scaling, so it is easier to use more samples to get a better model. Thus, BKRR2 is the most practical method because it balances the speed and the accuracy. BKRR3 can be used to evaluate the performance of systems both in accuracy and speed. KKRR2 is optimized for accuracy. We recommend using either BKRR2 or KKRR2.

## 5.5 Implementation Details

We use CSR (Compressed Row Storage) format for processing the sparse matrix (the original  $n$ -by- $d$  input matrix, not the dense kernel matrix) in our implementation. We use MPI for distributed processing. To give a fair comparison, we use ScaLAPACK [5] for solving the linear equation on distributed systems and LAPACK

[1] on shared memory systems, both of which are Intel MKL versions. Our experiments are conducted on NERSC Cori and Edison supercomputer systems [12]. Our source code is available online<sup>1</sup>.

The kernel matrix is symmetric positive definite (SPD) [17], and so is  $K + \lambda nI$ . Thus we use Cholesky decomposition to solve  $(K + \lambda nI)\alpha = y$  in DKRR, which is  $2.2\times$  faster than the Gaussian Elimination version.

To conduct a weak scaling study, we select a subset (e.g. 64k, 128k, 256k) from 463,715 samples as the training data to generate a model. Then we use the model to make a prediction. We use 2k test samples to show the MSE comparisons among different methods. If we change the number of test samples (e.g. change 2k to 20k), the testing accuracy will be roughly the same. The reason is that all the test samples are randomly shuffled so the 2k case will have roughly the sample pattern as the 20k case.

## 6 RELATED WORK

The most related work to this paper is DC-KRR [23]. The difference between DC-KRR and this work was presented in the Introduction section. Scholkopf et al [14] conducted a nonlinear form of principal component analysis (PCA) in high-dimensional feature spaces. Fine et al [6] used Incomplete Cholesky Factorization (ICF) to design an efficient Interior Point Method (IPM). Williams et al [18] designed a Nyström sampling method by carrying out an eigendecomposition on a smaller dataset. Si et al [16] designed a memory efficient kernel approximation method by first dividing the kernel matrix and then doing low-rank approximation on the smaller matrices. These methods can reduce the running time from  $\Theta(n^2d)$  to  $\Theta(nd^2)$  or  $\Theta(ndk)$  where  $k$  is the rank of the kernel approximation. However, all of these methods are proposed for serial analysis on single node systems. On the other hand, Kernel method is more accurate than existing approximate methods [23]. Zhang et al. [23] showed DC-KRR can beat all the previous approximate methods. DC-KRR is considered as state-of-the-art approach. Thus, we focus on the comparison with Zhang et al. [23].

<sup>1</sup><https://people.eecs.berkeley.edu/~youyang/cakrr.zip>

The second line of work is to use iterative methods such as gradient descent [20], block Jacobi method [15] and conjugate gradient method [4] to reduce the running time. However, iterative method can not make full use of computational powers efficiently because they only load a small amount of data to memory. If we scale the algorithm to 100+ cores, Kernel matrix method will be much faster. These methods provide a trade-off between time and accuracy, and we reserve them for future research. ASKIT [11] used n-body ideas to reduce the time and storage of kernel matrix evaluation, which is a direction of our future work. CA-SVM [21] [22] used the divide-and-conquer approach for machine learning applications. The differences include: (1) CA-SVM is for classification while this work is for regression. (2) CA-SVM uses an iterative method, whereas we use a direct method. (3) The iterative method did not store the huge Kernel matrix, thus CA-SVM work has no consideration on the huge Kernel matrix. The scalability of iterative method is also limited.

Lu et al [10] designed a fast ridge regression algorithm, which reduced the running time from  $\Theta(n^2d)$  to  $\Theta(\log(n)nd)$ . However, first, this work made an unreasonable assumption: the number of features is much larger than the number of samples ( $d \gg n$ ). This is true for some datasets like the webspam dataset [9], which has 350,000 training samples and each sample has 16,609,143 features. However, on average only about 10 of these 16 million features are nonzero. This can be processed efficiently by the sparse format like CSR (Compressed Sparse Row). Besides, this method only works for the linear ridge regression algorithm rather than the kernel version, which will lose non-linear space information.

## 7 CONCLUSION

Due to a  $\Theta(n^2)$  memory requirement and  $\Theta(n^3)$  arithmetic operations, KRR is prohibitive for large-scale machine learning datasets when  $n$  is very large. The weak scaling of KRR is problematic because the total memory required will grow like  $\Theta(p)$  per processor, and the total flops will grow like  $\Theta(p^2)$  per processor. The reason why BKRR2 is more scalable than DKRR is that it removes all the communication in the training part and reduces the computation cost from  $\Theta(n^3/p)$  to  $\Theta((n/p)^3)$ . The reason why BKRR2 is more accurate than DC-KRR is that it creates  $p$  different models and selects the best one. Compared to DKRR, BKRR2 improves the weak scaling efficiency from 0.32% to 92% and achieves 723 $\times$  speedup for the same accuracy on the same 1536 processors. Compared to DC-KRR, BKRR2 achieves much higher accuracy with faster speed. DC-KRR can never get the best accuracy achieved by BKRR2. When we increase # samples from 8k to 128k and # processors from 96 to 1536, BKRR2 reduces the MSE from 93 to 14.7, which solves the poor accuracy weak-scaling problem of DC-KRR (MSE only decreases from 89 to 81). KKRR2 achieves 591 $\times$  speedup over DKRR for the same accuracy by using the same data and hardware. Based on a variety of datasets used in this paper, we observe that KKRR2 is the most accurate method. BKRR2 is the most practical algorithm that balances the speed and accuracy. In conclusion, BKRR2 and KKRR2 are accurate, fast, and scalable methods.

## 8 ACKNOWLEDGEMENT

Yang You was supported by the U.S. DOE Office of Science, Office of Advanced Scientific Computing Research under Award Numbers DE-SC0008700 and AC02-05CH11231. Cho-Jui Hsieh acknowledges the support of NSF via IIS-1719097

## REFERENCES

- [1] Edward Anderson, Zhaojun Bai, Christian Bischof, Susan Blackford, Jack Dongarra, Jeremy Du Croz, Anne Greenbaum, Sven Hammarling, A McKeeney, and D Sorensen. 1999. *LAPACK Users' guide*. Vol. 9. Siam.
- [2] Ole E Barndorff-Nielsen and Neil Shephard. 2004. Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica* 72, 3 (2004), 885–925.
- [3] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset. In *ISMIR 2011: Proceedings of the 12th International Society for Music Information Retrieval Conference, October 24-28, 2011, Miami, Florida*. University of Miami, 591–596.
- [4] Gilles Blanchard and Nicole Krämer. 2010. Optimal learning rates for kernel conjugate gradient regression. In *Advances in Neural Information Processing Systems*. 226–234.
- [5] Jaeyoung Choi, James Demmel, Inderjit Dhillon, Jack Dongarra, Susan Ostrouchov, Antoine Petit, Ken Stanley, David Walker, and R Clinton Whaley. 1995. ScaLAPACK: A portable linear algebra library for distributed memory computers. In *Design issues and performance. In Applied Parallel Computing Computations in Physics, Chemistry and Engineering Science*. Springer, 95–106.
- [6] Shai Fine and Katya Scheinberg. 2002. Efficient SVM training using low-rank kernel representations. *The Journal of Machine Learning Research* 2 (2002), 243–264.
- [7] Edward W Forgy. 1965. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics* 21 (1965), 768–769.
- [8] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. 2008. Kernel methods in machine learning. *The annals of statistics* (2008), 1171–1220.
- [9] Chih-Jen Lin. 2017. LIBSVM Machine Learning Regression Repository. (2017). <https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>
- [10] Yichao Lu, Paramveer Dhillon, Dean P Foster, and Lyle Ungar. 2013. Faster ridge regression via the subsampled randomized hadamard transform. In *Advances in Neural Information Processing Systems*. 369–377.
- [11] William B March, Bo Xiao, and George Biros. 2015. ASKIT: Approximate skeletonization kernel-independent treecode in high dimensions. *SIAM Journal on Scientific Computing* 37, 2 (2015), A1089–A1110.
- [12] NERSC. 2016. NERSC Computational Systems. (2016). <https://www.nersc.gov/users/computational-systems/>
- [13] Mark Schmidt. 2005. Least squares optimization with l1-norm regularization. (2005).
- [14] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation* 10, 5 (1998), 1299–1319.
- [15] R. Schreiber. 1986. Solving eigenvalue and singular value problems on an undersized systolic array. *SIAM J. Sci. Stat. Comput.* 7 (1986), 441–451. first block Jacobi reference?
- [16] Si Si, Cho-Jui Hsieh, and Inderjit Dhillon. 2014. Memory efficient kernel approximation. In *Proceedings of The 31st International Conference on Machine Learning*. 701–709.
- [17] Leila Wehbe. 2013. Kernel Properties - Convexity. (2013).
- [18] Christopher Williams and Matthias Seeger. 2001. Using the Nystrom method to speed up kernel machines. In *Proceedings of the 14th Annual Conference on Neural Information Processing Systems*. 682–688.
- [19] Nicholas J WRIGHT, Sudip S DOSANJH, Allison K ANDREWS, Katerina B ANTYPAS, Brent DRANEY, R Shane CANON, Shreyas CHOLIA, Christopher S DALEY, Kirsten M FAGNAN, Richard A GERBER, et al. 2015. Cori: A Pre-Exascale Supercomputer for Big Data and HPC Applications. *Big Data and High Performance Computing* 26 (2015), 82.
- [20] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. 2007. On early stopping in gradient descent learning. *Constructive Approximation* 26, 2 (2007), 289–315.
- [21] Yang You, James Demmel, Kenneth Czechowski, Le Song, and Richard Vuduc. 2015. Ca-svm: Communication-avoiding support vector machines on distributed systems. In *Parallel and Distributed Processing Symposium (IPDPS), 2015 IEEE International*. IEEE, 847–859.
- [22] Yang You, Jim Demmel, Rich Vuduc, Le Song, and Kent Czechowski. 2016. Design and Implementation of a Communication-Optimal Classifier for Distributed Kernel Support Vector Machines. *IEEE Transactions on Parallel and Distributed Systems* (2016).
- [23] Yuchen Zhang, John Duchi, and Martin Wainwright. 2013. Divide and conquer kernel ridge regression. In *Conference on Learning Theory*. 592–617.