Distributed Nyström Kernel Learning with Communications

ICML 2021

Rong Yin, Yong Liu, Weiping Wang, and Dan Meng

Reporter: Rong Yin

Institute of Information Engineering, Chinese Academy of Sciences
yinrong@iie.ac.cn

June 21, 2021

Motivation

Given dataset $D = \cup_{j=1}^{p} D_j = \{(x_i, y_i)_{i=1}^{N}\}$ with p disjoint subsets $\{D_j\}_{j=1}^{p}$,
$\mathbf{y} = \mathbf{y}_D = [y_1, \ldots, y_{|D|}]^T$: data labels, $\mathbf{K}_N$: kernel matrix, $|D|$: the number of data in D.

Kernel Ridge Regression (KRR)

$$\hat{f}_{D,\lambda}(x) = \sum_{i=1}^{|D|} \hat{\alpha}_i K(x_i, x) \quad \text{with} \quad \hat{\boldsymbol{\alpha}} = (\mathbf{K}_N + \lambda |D| \mathbf{I})^{-1} \mathbf{y}, \tag{1}$$

are deduced from the square loss problem

$$\hat{f}_{D,\lambda} = \arg\min_{f \in \mathcal{H}} \frac{1}{|D|} \sum_{i=1}^{|D|} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2, \lambda > 0. \tag{2}$$

- Time complexity: $\mathcal{O}(|D|^3)$,
- Space complexity: $\mathcal{O}(|D|^2)$.

Contributions

In this paper, we study the statistical performance for distributed KRR with Nyström (DKRR-NY) and with Nyström and PCG (DKRR-NY-PCG).

Procedure: KRR $\longrightarrow \begin{cases} \text{KRR-NY} \\ \text{KRR-NY-PCG} \end{cases} \longrightarrow \begin{cases} \text{DKRR-NY} \longrightarrow \text{DKRR-NY-CM} \\ \text{DKRR-NY-PCG} \end{cases}$

$\Downarrow$

1. Our theoretical analysis show that DKRR-NY and DKRR-NY-PCG achieve the same optimal learning rates as the exact KRR requiring essentially $\mathcal{O}(|D|^{1.5})$ time and $\mathcal{O}(|D|)$ memory with relaxing the restriction on the number of local processors p in expectation, which exhibits the average effectiveness of multiple trials.

Note:

- DKRR-NY: distributed KRR with Nyström;
- DKRR-NY-PCG: distributed KRR with Nyström and PCG;
- DKRR-NY-CM: distributed KRR with Nyström and communication strategy.

Contributions

In this paper, we study the statistical performance for distributed KRR with Nyström (DKRR-NY) and with Nyström and PCG (DKRR-NY-PCG).

Procedure: KRR $\longrightarrow \left\{ \begin{array}{r} \text{KRR-NY} \\ \text{KRR-NY-PCG} \end{array} \right. \longrightarrow \left\{ \begin{array}{r} \text{DKRR-NY} \longrightarrow \text{DKRR-NY-CM} \\ \text{DKRR-NY-PCG} \end{array} \right.$

$\Downarrow$

1. Our theoretical analysis show that DKRR-NY and DKRR-NY-PCG achieve the same optimal learning rates as the exact KRR requiring essentially $\mathcal{O}(|D|^{1.5})$ time and $\mathcal{O}(|D|)$ memory with relaxing the restriction on the number of local processors p in expectation, which exhibits the average effectiveness of multiple trials.
2. For showing the generalization performance in a single trial, we deduce the optimal learning rates for DKRR-NY and DKRR-NY-PCG in probability.

Note:

- DKRR-NY: distributed KRR with Nyström;
- DKRR-NY-PCG: distributed KRR with Nyström and PCG;
- DKRR-NY-CM: distributed KRR with Nyström and communication strategy.

Contributions

In this paper, we study the statistical performance for distributed KRR with Nyström (DKRR-NY) and with Nyström and PCG (DKRR-NY-PCG).

Procedure: KRR $\longrightarrow$ { KRR-NY ; KRR-NY-PCG } $\longrightarrow$ { DKRR-NY $\longrightarrow$ DKRR-NY-CM ; DKRR-NY-PCG }

$\Downarrow$

1. Our theoretical analysis show that DKRR-NY and DKRR-NY-PCG achieve the same **optimal learning rates** as the exact KRR requiring essentially $\mathcal{O}(|D|^{1.5})$ **time and** $\mathcal{O}(|D|)$ **memory with relaxing the restriction on the number of local processors** p **in expectation**, which exhibits the average effectiveness of multiple trials.
2. For showing the generalization performance in a single trial, we deduce the **optimal learning rates** for DKRR-NY and DKRR-NY-PCG **in probability.**
3. We propose a novel algorithm DKRR-NY-CM based on DKRR-NY, which employs a communication strategy to further improve the learning performance, whose effectiveness of communications is validated in theoretical and experimental analysis.

Note:

- DKRR-NY: distributed KRR with Nyström;
- DKRR-NY-PCG: distributed KRR with Nyström and PCG;
- DKRR-NY-CM: distributed KRR with Nyström and communication strategy.

KRR with Nyström (KRR-NY)

Consider a smaller hypothesis space $\mathcal{H}_m$

$$\mathcal{H}_m = \{f | f = \sum_{i=1}^{m} \alpha_i K(\tilde{x}_i, \cdot), \boldsymbol{\alpha} \in \mathbb{R}^m\}$$

of functions

$$\tilde{f}_{m,\lambda}(x) = \sum_{i=1}^{m} \tilde{\alpha}_i K(\tilde{x}_i, x), \quad (3)$$

The corresponding minimizer over the space $\mathcal{H}_m$ is

$$\tilde{\boldsymbol{\alpha}} = \underbrace{(\mathbf{K}_{Nm}^T \mathbf{K}_{Nm} + \lambda |D| \mathbf{K}_{mm})}_{\mathbf{H}}^{\dagger} \underbrace{\mathbf{K}_{Nm}^T \mathbf{y}}_{\mathbf{z}}. \quad (4)$$

where $\{\tilde{x}_1, \ldots, \tilde{x}_m\}$ are Nyström centers sampled uniformly at random without replacement from the training set.

Procedure: KRR $\longrightarrow$ { KRR-NY / KRR-NY-PCG } $\longrightarrow$ { DKRR-NY $\longrightarrow$ DKRR-NY-CM / DKRR-NY-PCG }

To quickly compute $\tilde{\alpha}$ in the above (Eq.(4)), preconditioning and conjugate gradient (PCG) is introduced.

KRR with Nyström and PCG (KRR-NY-PCG)

$$\mathbf{P}^T\mathbf{H}\hat{\alpha} = \mathbf{P}^T\mathbf{z}, \quad \text{with} \quad \hat{f}_{m,\lambda}(x) = \sum_{i=1}^{m} \hat{\alpha}_i K(\tilde{x}_i, x), \tag{5}$$

where

- $\hat{\alpha}$ is solved via t-step conjugate gradient algorithm,
- $\mathbf{P} = \frac{1}{\sqrt{|D|}}\mathbf{T}^{-1}\mathbf{A}^{-1}$,
- $\mathbf{T} = \text{chol}(\mathbf{K}_{mm})$,
- $\mathbf{A} = \text{chol}(\frac{1}{m}\mathbf{T}\mathbf{T}^T + \lambda\mathbf{I})$,
- chol() represents the Cholesky decomposition.

Procedure: KRR $\longrightarrow$ $\begin{cases} \text{KRR-NY} \\ \text{KRR-NY-PCG} \end{cases}$ $\longrightarrow$ $\begin{cases} \text{DKRR-NY} \longrightarrow \text{DKRR-NY-CM} \\ \text{DKRR-NY-PCG} \end{cases}$

DKRR-NY-PCG

$$\bar{f}^{0}_{D,m,t} = \sum_{j=1}^{p} \frac{|D_j|}{|D|} f_{D_j,m,t}, \tag{6}$$

where $f_{D_j,m,t}$ is the solver of KRR-NY-PCG in Eq.(5).
When $t \to \infty$, Eq.(6) is distributed KRR-NY (DKRR-NY), $f_{D_j,m,t}$ is rewritten as $f_{D_j,m,\lambda}$.

Procedure: KRR $\longrightarrow$ { **KRR-NY** / **KRR-NY-PCG** } $\longrightarrow$ { **DKRR-NY** $\longrightarrow$ DKRR-NY-CM / **DKRR-NY-PCG** }

Theorem (DKRR-NY in Expectation)

Under basic Assumptions, let $r \in [1/2, 1]$, $\gamma \in (0, 1]$, $\lambda = \Omega(|D|^{-\frac{1}{2r+\gamma}})$, *with probability* $1 - \delta$, *when* $p \leq \mathcal{O}(|D|^{\frac{2r+\gamma-1}{2r+\gamma}})$ *and* $m \geq \mathcal{O}(|D|^{\frac{1}{2r+\gamma}})$, *we have*
$\mathbb{E}[\mathcal{E}(\bar{f}^0_{D,m,\lambda})] - \mathcal{E}(f_{\mathcal{H}}) = \mathcal{O}(|D|^{-\frac{2r}{2r+\gamma}})$.

Corollary (DKRR-NY-PCG in Expectation)

Under basic Assumptions, let $r \in [1/2, 1]$, $\gamma \in (0, 1]$, $\lambda = \Omega(|D|^{-\frac{1}{2r+\gamma}})$, *with probability* $1 - \delta$, *when* $t \geq \mathcal{O}(\log(|D|))$, $p \leq \mathcal{O}(|D|^{\frac{2r+\gamma-1}{2r+\gamma}})$, *and* $m \geq \mathcal{O}(|D|^{\frac{1}{2r+\gamma}})$, *we have*
$\mathbb{E}[\mathcal{E}(\bar{f}^0_{D,m,t})] - \mathcal{E}(f_{\mathcal{H}}) = \mathcal{O}(|D|^{-\frac{2r}{2r+\gamma}})$.

NOTE:

- $\mathcal{O}(N^{-\frac{2r}{2r+\gamma}})$ is the optimal learning rate of KRR.
- Under the basic setting ($r = 1/2$ and $\gamma = 1$), the upper bound of the number of local processors p is enlarged from $\mathcal{O}(1)$ of previous work [Yin et al., 2020a] to our $\mathcal{O}(\sqrt{|D|})$ with the optimal learning rate.

Procedure: KRR $\longrightarrow$ { KRR-NY / KRR-NY-PCG } $\longrightarrow$ { DKRR-NY $\longrightarrow$ DKRR-NY-CM / DKRR-NY-PCG }

Theoretical Analysis of DKRR-NY and DKRR-NY-PCG in Probability

For showing the generalization performance in a single trial, we deduce the learning rates for DKRR-NY and DKRR-NY-PCG in probability.

Theorem (DKRR-NY in Probability)

Under basic Assumptions, let $r \in [1/2, 1]$, $\gamma \in (0, 1]$, $\lambda = \Omega(|D|^{-\frac{1}{2r+\gamma}})$, *with probability* $1-\delta$, *when* $p \leq \mathcal{O}(|D|^{\frac{2r+\gamma-1}{4r+2\gamma}})$ *and* $m \geq \mathcal{O}(|D|^{\frac{1}{2r+\gamma}})$, *we have* $\|\bar{f}^0_{D,m,\lambda} - f_{\mathcal{H}}\|^2_\rho = \mathcal{O}(|D|^{-\frac{2r}{2r+\gamma}})$.

Corollary (DKRR-NY-PCG in Probability)

Under basic Assumptions, let $r \in [1/2, 1]$, $\gamma \in (0, 1]$, $\lambda = \Omega(|D|^{-\frac{1}{2r+\gamma}})$, *with probability* $1-\delta$, *when* $t \geq \mathcal{O}(\log(|D|))$, $p \leq \mathcal{O}(|D|^{\frac{2r+\gamma-1}{4r+2\gamma}})$, *and* $m \geq \mathcal{O}(|D|^{\frac{1}{2r+\gamma}})$, *we have* $\|\bar{f}^0_{D,m,t} - f_{\mathcal{H}}\|^2_\rho = \mathcal{O}(|D|^{-\frac{2r}{2r+\gamma}})$.

Note:

- Since the error decomposition in probability is not easy to separate a distributed error to control the number of local processors, the upper bound $\mathcal{O}(|D|^{\frac{2r+\gamma-1}{4r+2\gamma}})$ of p in probability is stricter than $\mathcal{O}(|D|^{\frac{2r+\gamma-1}{2r+\gamma}})$ in expectation.

Procedure: KRR $\longrightarrow$ { KRR-NY / KRR-NY-PCG } $\longrightarrow$ { DKRR-NY $\longrightarrow$ DKRR-NY-CM / DKRR-NY-PCG }

To further enlarge the number of local processors p, we present a novel communication strategy for DKRR-NY (called DKRR-NY-CM).

DKRR-NY-CM

$$\bar{f}^{l}_{D,m,\lambda} = \bar{f}^{l-1}_{D,m,\lambda} - \sum_{j=1}^{p} \frac{|D_j|}{|D|} \beta_j^{l-1}, l > 0$$

where
$\beta_j^{l-1} = (P_m C_n P_m + \lambda I)^{-1} G_{D,m,\lambda}(\bar{f}^{l-1}_{D,m,\lambda})$,
local gradient: $G_{D_j,m,\lambda}(f) = (P_m C_n P_m + \lambda I) f - \frac{1}{\sqrt{|D_j|}} P_m S_n^* \mathbf{y}_{D_j}$, and
global gradient: $G_{D,m,\lambda}(f) = \sum_{j=1}^{p} \frac{|D_j|}{|D|} G_{D_j,m,\lambda}(f)$.

Note:

- DKRR-NY-CM communicates the gradients instead of data between local processors, which can protect the privacy of datasets in each local processor.

Procedure: KRR $\longrightarrow \begin{cases} \text{KRR-NY} \\ \text{KRR-NY-PCG} \end{cases} \longrightarrow \begin{cases} \text{DKRR-NY} \longrightarrow \text{DKRR-NY-CM} \\ \text{DKRR-NY-PCG} \end{cases}$

Theorem (3DKRR-NY-CM in Probability)

Under basic Assumptions, let $r \in [1/2, 1]$, $\gamma \in (0, 1]$, $\lambda = \Omega(|D|^{-\frac{1}{2r+\gamma}})$, *with probability* $1-\delta$, *when* $p \leq \mathcal{O}(|D|^{\frac{(2r+\gamma-1)(M+1)}{(2r+\gamma)(M+2)}})$ *and* $m \geq \mathcal{O}(|D|^{\frac{1}{2r+\gamma}})$, *we have*
$\|\bar{f}^M_{D,m,\lambda} - f_{\mathcal{H}}\|^2_\rho = \mathcal{O}(|D|^{-\frac{2r}{2r+\gamma}})$.

Note:

- DKRR-NY-CM enlarges the upper bound of p compared with DKRR-NY: $p \leq \mathcal{O}(|D|^{\frac{2r+\gamma-1}{4r+2\gamma}})$ —> $p \leq \mathcal{O}(|D|^{\frac{(2r+\gamma-1)(M+1)}{(2r+\gamma)(M+2)}})$.
- The upper bound of p is monotonically increasing with the number of communications M, showing the power of communications.

Procedure: KRR $\longrightarrow$ {KRR-NY, KRR-NY-PCG} $\longrightarrow$ {DKRR-NY $\longrightarrow$ DKRR-NY-CM, DKRR-NY-PCG}

Compared Methods

Table 1: Computational complexity of the approximation KRR with the optimal learning rate and $\lambda = 1/\sqrt{|D|}$. "Comm" is communication complexity. $d > 0$, $\Delta_1 = \frac{(1-\gamma)\gamma}{2} \geq 0$, $\Delta_2 = \frac{\gamma}{2} > 0$, and $\gamma \in (0, 1]$.

Algorithms	Time	Space	Comm	p	m	Types
Nyström[Rudi et al., 2015]	$\lvert D\rvert^2$	$\lvert D\rvert^{1.5}$	/	/	$\lvert D\rvert^{0.5}$	In probability
Nyström-PCG[Rudi et al., 2017]	$\lvert D\rvert^{1.5}$	$\lvert D\rvert^{1.5}$	/	/	$\lvert D\rvert^{0.5}$	In probability
Random Features[Rudi et al., 2016]	$\lvert D\rvert^{2+2\Delta_1}$	$\lvert D\rvert^{1.5+\Delta_1}$	/	/	$\lvert D\rvert^{0.5+\Delta_1}$	In probability
DKRR-RF[Li et al., 2019]	$\lvert D\rvert^{1.5+2\Delta_1+\Delta_2}$	$\lvert D\rvert^{1+\Delta_1+\Delta_2}$	$\lvert D\rvert^{0.5+\Delta_1}$	$\lvert D\rvert^{0.5-\Delta_2}$	$\lvert D\rvert^{0.5+\Delta_1}$	In expectation
DKRR-RF[Liu et al., 2021]	$\lvert D\rvert^{1.5+2\Delta_1}$	$\lvert D\rvert^{1+\Delta_1}$	$\lvert D\rvert^{0.5+\Delta_1}$	$\lvert D\rvert^{0.5}$	$\lvert D\rvert^{0.5+\Delta_1}$	In expectation
DKRR-RF[Liu et al., 2021]	$\lvert D\rvert^{1.75+2\Delta_1}$	$\lvert D\rvert^{1.25+\Delta_1}$	$\lvert D\rvert^{0.5+\Delta_1}$	$\lvert D\rvert^{0.25}$	$\lvert D\rvert^{0.5+\Delta_1}$	In probability
DKRR-RF-CM[Liu et al., 2021]	$\lvert D\rvert^{\frac{3M+7}{2M+4}+2\Delta_1}$	$\lvert D\rvert^{\frac{2M+5}{2M+4}+\Delta_1}$	$M\lvert D\rvert^{0.5+\Delta_1}$	$\lvert D\rvert^{\frac{M+1}{2(M+2)}}$	$\lvert D\rvert^{0.5+\Delta_1}$	In probability
DKRR[Chang et al., 2017b]	$\lvert D\rvert^2$	$\lvert D\rvert$	$\lvert D\rvert^{0.5}$	$\lvert D\rvert^{0.5}$	/	In expectation
DKRR[Lin et al., 2020]	$\lvert D\rvert^{2.25}$	$\lvert D\rvert^{1.5}$	$\lvert D\rvert^{0.75}$	$\lvert D\rvert^{0.25}$	/	In probability
DKRR-CM[Lin et al., 2020]	$\lvert D\rvert^{\frac{3(M+3)}{2(M+2)}}$	$\lvert D\rvert^{\frac{M+3}{M+2}}$	$Md\lvert D\rvert$	$\lvert D\rvert^{\frac{M+1}{2(M+2)}}$	/	In probability
DKRR-NY-PCG[Yin et al., 2020a]	$\lvert D\rvert^{1.5}$	$\lvert D\rvert^{1+\Delta_2}$	$\lvert D\rvert^{0.5}$	$\lvert D\rvert^{0.5-\Delta_2}$	$\lvert D\rvert^{0.5}$	In expectation
DKRR-NY-PCG [This paper]	$\lvert D\rvert^{1.5}$	$\lvert D\rvert$	$\lvert D\rvert^{0.5}$	$\lvert D\rvert^{0.5}$	$\lvert D\rvert^{0.5}$	In expectation
DKRR-NY-PCG [This paper]	$\lvert D\rvert^{1.75}$	$\lvert D\rvert^{1.25}$	$\lvert D\rvert^{0.5}$	$\lvert D\rvert^{0.25}$	$\lvert D\rvert^{0.5}$	In probability
DKRR-NY [This paper]	$\lvert D\rvert^{1.5}$	$\lvert D\rvert$	$\lvert D\rvert^{0.5}$	$\lvert D\rvert^{0.5}$	$\lvert D\rvert^{0.5}$	In expectation
DKRR-NY [This paper]	$\lvert D\rvert^{1.75}$	$\lvert D\rvert^{1.25}$	$\lvert D\rvert^{0.5}$	$\lvert D\rvert^{0.25}$	$\lvert D\rvert^{0.5}$	In probability
DKRR-NY-CM [This paper]	$\lvert D\rvert^{\frac{3M+7}{2M+4}}$	$\lvert D\rvert^{\frac{2M+5}{2M+4}}$	$M\lvert D\rvert^{0.5}$	$\lvert D\rvert^{\frac{M+1}{2(M+2)}}$	$\lvert D\rvert^{0.5}$	In probability

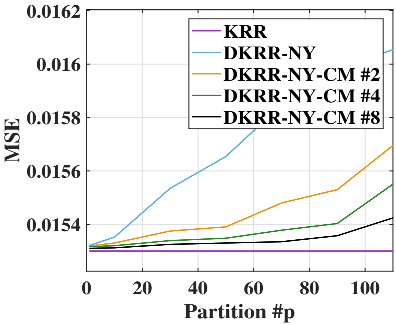

Figure 1: The mean square error on testing sampling with different partitions on KRR, DKRR-NY, and our DKRR-NY-CM. The numbers 2, 4 and 8 represent the number of communications.

Thank You for Listening