Distributed Generalized Cross-Validation for Divide-and-Conquer Kernel Ridge Regression and its Asymptotic Optimality

Ganggang Xu[1], Zuofeng Shang[2], and Guang Cheng[3] [4]

February 19, 2019

Abstract

Tuning parameter selection is of critical importance for kernel ridge regression. To this date, data driven tuning method for divide-and-conquer kernel ridge regression (d-KRR) has been lacking in the literature, which limits the applicability of d-KRR for large data sets. In this paper, by modifying the Generalized Cross-validation (GCV, Wahba, 1990) score, we propose a distributed Generalized Cross-Validation (dGCV) as a data-driven tool for selecting the tuning parameters in d-KRR. Not only the proposed dGCV is computationally scalable for massive data sets, it is also shown, under mild conditions, to be asymptotically optimal in the sense that minimizing the dGCV score is equivalent to minimizing the true global conditional empirical loss of the averaged function estimator, extending the existing optimality results of GCV to the divide-and-conquer framework.

1 Introduction

Massive data made available in various research areas have imposed new challenges for data scientists. With a large to massive sample size, many sophisticated statistical tools are no longer applicable simply due to formidable computational costs and/or memory

[1]Department of Management Science, University of Miami, Coral Gables, 33146, USA.
E-mail: gangxu@bus.miami.edu

[2]Department of Mathematical Sciences, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202, USA.
E-mail: shangzf@iu.edu

[3]Department of Statistics, Purdue University, West Lafayette, IN 47907-2066 USA.
E-mail: chengg@stat.purdue.edu

[4]This paper is an extended version of the work published in the conference paper Xu et al. (2018).

requirements. Even when the computation is possible on more advanced machines, it is still appealing to develop accurate statistical procedures at much lower computational costs. The divide-and-conquer strategy has become a popular tool for regression models. With carefully designed algorithms, such a strategy has proven to be effective in Linear models (Chen and Xie, 2014; Lu et al., 2016), Partially linear models (Zhao et al., 2016) and Nonparametric regression models (Zhang et al., 2015; Lin et al., 2017; Shang and Cheng, 2017; Guo et al., 2017). In this paper, we shall focus on the divide-and-conquer kernel ridge regression (d-KRR) where the selection of the penalty parameter is of vital importance but still remains unsettled.

Suppose we have independent and identically distributed samples $\{(x_i, y_i) \in \mathcal{X} \times \mathbb{R}\}_{i=1,\dots N}$ from a joint probability measure $\mathbb{P}_{Y,X}$. The goal is to study the association between the covariate vector $\boldsymbol{x}_i$ and the response y_i through the following model

$$y_i = f_0(\boldsymbol{x}_i) + \varepsilon_i, \quad i = 1, \dots, N, \tag{1}$$

where $f_0(\cdot) : \mathcal{X} \to \mathbb{R}$ is the function of interest and ε_i is a random error term with mean zero and a common variance σ^2. One popular method to estimate $f_0(\cdot)$ is the *Kernel Ridge Regression* (Shawe-Taylor and Cristianini, 2004) which essentially aims at finding a projection of $f_0(\cdot)$ into a reproducing kernel Hilbert space (RKHS), denoted as $\mathcal{H}$, equipped with a norm $\|\cdot\|_{\mathcal{H}}$. Specifically, the KRR estimator is then defined as

$$\widehat{f} = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - f(\boldsymbol{x}_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}, \tag{2}$$

where $\lambda \geq 0$ controls trade-off between goodness-of-fit and smoothness of f.

It is well known that computing $\widehat{f}$ requires $O(N^3)$ floating operations and $O(N^2)$ memory; see (5) for more details. When N is large, such requirements can be prohibitive. To overcome this, Zhang et al. (2015) proposed the following "divide-and-conquer" algorithm: (i) Randomly divide the entire sample $\{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_N, y_N)\}$ to m disjoint "smaller" subsets, denoted by $S_1, \dots, S_m$; (ii) For each subset S_k, find $\widehat{f}_k = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n_k} \sum_{i \in S_k} (y_i - f(\boldsymbol{x}_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}$, where n_k is the size of S_k; (iii) The final nonparametric estimator is given by

$$\bar{f}(x) = \frac{1}{m} \sum_{k=1}^{m} \widehat{f}_k(\boldsymbol{x}). \tag{3}$$

Such a "divide-and-conquer" strategy reduces computing time from $O(N^3)$ to $O(N^3/m^2)$ and memory usage from $O(N^2)$ to $O(N^2/m^2)$. Both savings may be substantial as m grows. Furthermore, Zhang et al. (2015) shows that as long as m does not grow too fast, the averaged estimator $\bar{f}$ achieves the same minimax optimal estimation rate as the oracle estimate $\widehat{f}$, i.e., (2), that utilizes all data points at once. In this sense, the divide-and-conquer algorithm is quite appealing as it achieves an ideal balance between

the computational cost and the statistical efficiency.

However, the aforementioned statistical efficiency depends critically on a careful choice of tuning parameter λ in all sub-samples. The optimal choice of tuning parameter λ has been well studied for KRR when the entire data set can be fitted at once. Examples include Mallow's CP (Mallows, 2000), Generalized cross-validation (GCV, Craven and Wahba, 1978) and Generalized approximated cross-validation (Xiang and Wahba, 1996). However, if we naively apply these traditional tuning methods in each sub-sample to pick an optimal λ_k in the above step (ii), the averaged function estimator $\bar{f}$ subsequently obtained using (3) will be sub-optimal. As pointed out by existing literature (e.g. Zhang et al., 2015; Blanchard and Mücke, 2016; Chang et al., 2017), the optimal tuning parameter should be chosen in accordance with the order of *the entire sample size*, i.e., N, such that we intentionally allow the resulting sub-estimator $\widehat{f}_k$ to over-fit the sub-sample S_k for each $k = 1, \ldots, m$. Based on the order of the optimal choice of λ, Zhang et al. (2015) proposed a heuristic data-driven approach to empirically choose an optimal λ. However, the theoretical properties of this approach remain unclear. In this paper, we define a new data-driven criterion named "distributed generalized cross-validation" (dGCV) to choose tuning parameters for KRR under the divide-and-conquer framework. The computational cost of the proposed criterion remains the same as $O(N^3/m^2)$. More importantly, we show that the proposed method enjoys similar theoretical optimality as the well-known GCV criterion (Craven and Wahba, 1978) in the sense that the resulting divide-and-conquer estimate minimizes the true empirical loss function asymptotically.

The rest of paper are organized as follows. Section 2 introduces background on kernel ridge regression. Section 3 presents the main result of this paper on the dGCV, while Section 4 gives statistical guarantee for this new tuning procedure. Our method and theory are backed up by extensive simulation studies in Sections 5, and are applied to the Million Song Dataset in Section 6, demonstrating significant advantages over Zhang et al. (2015). All technical proofs are postponed to the Appendix.

2 Kernel Ridge Regression Estimation

In this section, we briefly review kernel ridge regression (Shawe-Taylor and Cristianini, 2004). The reproducing kernel Hilbert space, denoted as $\mathcal{H}$, is a Hilbert space induced by a symmetric nonnegative definite kernel function $K(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ satisfying

$$\langle g(\cdot), K(\boldsymbol{x}, \cdot) \rangle_{\mathcal{H}} = g(\boldsymbol{x}) \text{ for any } g \in \mathcal{H}.$$

The kernel function $K(\cdot, \cdot)$ is called the reproducing kernel of the Hilbert space $\mathcal{H}$ equipped with the norm $\|g\|_{\mathcal{H}} = \sqrt{\langle g(\cdot), g(\cdot) \rangle_{\mathcal{H}}}$. Using the Mercer's theorem, under some regularity conditions, the kernel function $K(\cdot, \cdot)$ possesses the expansion $K(\boldsymbol{x}, \boldsymbol{z}) =$

$\sum_{j=1}^{\infty} \mu_j \psi_j(\boldsymbol{x})\psi_j(\boldsymbol{z})$, where $\mu_1 \geq \mu_2 \geq \dots$ is a sequence of decreasing eigenvalues and $\{\psi_1(\cdot), \psi_2(\cdot), \dots\}$ is a family of orthonormal basis functions of $L^2(\mathbb{P}_X)$. The smoothness of $g \in \mathcal{H}$ is characterized by the decaying rate of the eigenvalues $\{\mu_j\}_{j=1}^{\infty}$. There are three types of estimation considered in this paper, including smoothing spline (Wahba, 1990) as a special case.

Finite rank: There exists some integer r such that $\mu_j = 0$ for $j > r$. For example, with vectors $\boldsymbol{x}, \boldsymbol{z}$, the polynomial kernel $K(\boldsymbol{x}, \boldsymbol{z}) = (1 + \boldsymbol{x}^T\boldsymbol{z})^r$ has a finite rank $r + 1$, and induces a space of polynomial functions with degree at most r. This corresponds to the parametric ridge regression.

Exponentially decaying: There exist some $\alpha, r > 0$ such that $\mu_j \asymp \exp(-\alpha j^r)$. Exponentially decaying kernels include the multivariate Gaussian kernel $K(\boldsymbol{x}, \boldsymbol{z}) = \exp(-\|\boldsymbol{x} - \boldsymbol{z}\|_2^2/\phi^2)$, where $\phi > 0$ is the scale parameter and $\|\cdot\|_2$ is the Euclidean norm.

Polynomially decaying: There exists some $r > 0$ such that $\mu_j \asymp j^{-2r}$. The polynomially decaying class includes many smoothing spline kernels of the Sobolev space (Wahba, 1990). For example, kernel function $K(x, z) = 1+\min(x, z)$ induces the Sobolev space of Lipschitz functions with smoothness $\nu = 1$ and has polynomially decaying eigenvalues.

2.1 The Representer Theorem

With observed data, using the representer theorem (Wahba, 1990), it can be shown that the solution to the minimization problem (2) takes the following form

$$\widehat{f}(\boldsymbol{x}) = \sum_{i=1}^{N} \beta_i K(\boldsymbol{x}_i, \boldsymbol{x}), \tag{4}$$

where $\beta_1, \dots, \beta_N \in \mathbb{R}$. Furthermore, based on the observed sample, the parameter vector $\boldsymbol{\beta} = (\beta_1, \dots \beta_N)^T$ can be estimated by minimizing the following criterion

$$\frac{1}{N}(\mathbf{Y} - \boldsymbol{\beta}^T\mathbf{K})^T(\mathbf{Y} - \boldsymbol{\beta}^T\mathbf{K}) + \lambda\boldsymbol{\beta}^T\mathbf{K}\boldsymbol{\beta}, \tag{5}$$

where $\mathbf{Y} = (y_1, \dots, y_N)^T$ and $\mathbf{K} = [K(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1,\dots,N}$. The solution to (5) takes the form of $\widehat{\boldsymbol{\beta}} = (\mathbf{K} + N\lambda\mathbf{I}_N)^{-1}\mathbf{Y}$, which requires $O(N^3)$ operations.

We next apply the above idea to sub-estimation. Denote $(\mathbf{y}_1, \mathbf{x}_1), \dots, (\mathbf{y}_m, \mathbf{x}_m)$ as a random partition of the entire data with $\mathbf{y}_k = (y_{k,1}, \dots, y_{k,n_k})^T$ and $\mathbf{x}_k = (\boldsymbol{x}_{k,1}, \dots, \boldsymbol{x}_{k,n_k})^T$. Define vectors $\mathbf{f}_k = (f_0(\boldsymbol{x}_{k,1}), \dots, f_0(\boldsymbol{x}_{k,n_k}))^T$ and $\boldsymbol{\varepsilon}_k = \mathbf{y}_k - \mathbf{f}_k$. Define the sub-kernel matrices $\mathbf{K}_{kl} = [K(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i\in S_k, j\in S_l}$ for $l, k = 1, \dots, m$. It is straightforward to show that the minimizer of (5) with $\mathbf{K}$ replaced by $\mathbf{K}_{kk}$ is of the form

$\widehat{\boldsymbol{\beta}}_k = (\mathbf{K}_{kk} + n_k\lambda\mathbf{I}_k)^{-1}\mathbf{y}_k$, and the individual function estimator $\widehat{f}_k(x)$ can be written as

$$\widehat{f}_k(\boldsymbol{x}) = \sum_{i\in S_k} \widehat{\beta}_{k,i} K(\boldsymbol{x}_i, \boldsymbol{x}), \tag{6}$$

where $\widehat{\beta}_{k,i}$ is the entry of $\widehat{\boldsymbol{\beta}}_k$ corresponding to $x_{k,i}$, $k = 1, \ldots, m$.

2.2 Kernel Ridge Regression for Multivariate Functions

In principle, any multivariate function $f_0(\boldsymbol{x})$ in (1), i.e., $\boldsymbol{x} \in \mathbb{R}^p$, can be well approximated if a sufficiently good reproducing kernel $K(\cdot,\cdot)$ can be identified. However, for a large p, the excessive risk of the KRR estimator may grow exponentially fast as the dimension p increases (Györfi et al., 2006), which is often referred to as the "curse of dimensionality". One common strategy is to impose some special structures on the reproducing kernel. For example, the polynomial kernel $K(\boldsymbol{x}, \boldsymbol{z}) = (1 + \boldsymbol{x}^T\boldsymbol{z})^r$ assumes that $K(\cdot,\cdot)$ depends only on the inner product of $\boldsymbol{x}$ and $\boldsymbol{z}$ and the multivariate Gaussian kernel $K(\boldsymbol{x}, \boldsymbol{z}) = \exp(-\|\boldsymbol{x} - \boldsymbol{z}\|_2^2/\phi^2)$ assumes that $K(\cdot,\cdot)$ is determined by the Euclidean distance between vectors $\boldsymbol{x}$ and $\boldsymbol{z}$. More sophisticated applications of Gaussian kernels may also allow the scale parameter ϕ to vary for different dimensions. Another popular approach to circumvent the "curse of dimensionality" is to use additive approximation (Hastie, 2017; Kandasamy and Yu, 2016) to multivariate functions. Let $\boldsymbol{x} = (x_1, x_2, \cdot, x_p)^T$, and define the first-order additive approximation of $f(\boldsymbol{x})$ as

$$f^*(\boldsymbol{x}) = f_1^*(x_1) + \cdots + f_p^*(x_p), \tag{7}$$

where each $f_j^*(\cdot)$ is a univariate function residing in a reproducing kernel Hilbert space $\mathcal{H}_k$ with a reproducing kernel $K_j(\cdot,\cdot)$, $j = 1, \cdots, p$. The corresponding additive kernel can be defined as $K(\boldsymbol{x}, \boldsymbol{z}) = \sum_{j=1}^p K_j(x_j, z_j)$, and the associated reproducing kernel Hilbert space is $\mathcal{H} = \mathcal{H}_1 \bigoplus \mathcal{H}_2 \bigoplus \cdots \bigoplus \mathcal{H}_p$. For some applications where the first order approximation (7) is not adequate, higher order additive approximations to the multivariate function $f(\boldsymbol{x})$ can be used to achieve better estimation accuracies at similar computational costs, see Kandasamy and Yu (2016) for more detailed discussions.

3 Tuning Parameter Selection

3.1 Sub-GCV Score: Local Optimality

In this section, we define the GCV score for each sub-estimation, named as sub-GCV score, and discuss its theoretical property. Define the empirical loss function for $\widehat{f}_k$ as follows

$$L_k(\lambda|\mathbf{x}_k) = \frac{1}{n_k}\sum_{i\in S_k} w_i \left\{\widehat{f}_k(x_i) - f_0(x_i)\right\}^2, \tag{8}$$

where $w_i \geq 0$ is some weight assigned to each observation (y_i, x_i) and satisfies $\sum_{i \in S_k} w_i = n_k$. The introduction of weights in (8) helps reducing computational cost; see Section 3.4. The tuning parameter λ is referred to as "locally optimal" if it only minimizes local empirical loss $L_k(\lambda|\mathbf{x}_k)$. When only focused on a single sub-data set, such a "locally-optimal" choice of tuning parameter λ has been well studied in (Craven and Wahba, 1978; Li, 1986; Gu, 2013; Wood, 2004; Gu and Ma, 2005; Xu and Huang, 2012), among which the Generalized Cross-Validation (Craven and Wahba, 1978) remains to be one of the most popular approaches.

Using the function estimator $\widehat{f}_k(x)$, the predicted values for the vector $\mathbf{y}_k$ can be written as $\widehat{\mathbf{y}}_k = \mathbf{A}_{kk}(\lambda)\mathbf{y}_k$, where $\mathbf{A}_{kk}(\lambda) = \mathbf{K}_{kk}(\mathbf{K}_{kk} + n_k\lambda\mathbf{I}_k)^{-1}$. Here the matrix $\mathbf{A}_{kk}(\lambda)$ is often known as the hat matrix. Using the above notations, the sub-GCV score is defined as

$$\mathrm{GCV}_k(\lambda) = \frac{n_k^{-1}(\widehat{\mathbf{y}}_k - \mathbf{y}_k)^T\mathbf{W}_k(\widehat{\mathbf{y}}_k - \mathbf{y}_k)}{\{1 + n_k^{-1}\mathrm{tr}\{\mathbf{A}_{kk}(\lambda)\mathbf{W}_k\}\}^2}, \tag{9}$$

where $\mathbf{W}_k = \mathrm{diag}\{w_i, i \in S_k\}$, $k = 1, \ldots, m$. It is well known that $\mathrm{GCV}_k(\lambda)$ enjoys appealing asymptotic properties. For example, under mild conditions, Gu (2013) showed that, as $n_k \to \infty$,

$$\mathrm{GCV}_k(\lambda) - L_k(\lambda|\mathbf{x}_k) - \frac{1}{n_k}\boldsymbol{\varepsilon}_k^T\mathbf{W}_k\boldsymbol{\varepsilon}_k = o_{\mathbb{P}_\varepsilon}\{L_k(\lambda|\mathbf{x}_k)\},$$

$k = 1, \ldots, m$. This property essentially asserts that, minimizing $\mathrm{GCV}_k(\lambda)$ with respect to λ is asymptotically equivalently to minimizing the local "golden criterion" $L_k(\lambda|\mathbf{x}_k)$.

3.2 Local-Optimality v.s. Global-Optimality

In this section, we explain why the use of $\mathrm{GCV}_k(\lambda)$ in each subsample does not lead to an optimal averaged estimate $\bar{f}$. We first derive conditional risks for both $\widehat{f}_k$ and $\bar{f}$. For the former, some basic algebra yields that the conditional risk $R_k(\lambda|\mathbf{x}_k) = \mathbb{E}_\varepsilon\{L_k(\lambda|\mathbf{x}_k)\}$ is of the form

$$R_k(\lambda|\mathbf{x}_k) = \frac{1}{n_k}\sum_{i \in S_k} w_i \mathrm{Var}_\varepsilon\left\{\widehat{f}_k(x_i)\right\} + \frac{1}{n_k}\sum_{i \in S_k} w_i\left\{\mathbb{E}_\varepsilon\widehat{f}_k(x_i) - f_0(x_i)\right\}^2, \tag{10}$$

where the expectation is taken with respect to the probability measure $\mathbb{P}_\varepsilon$. As for the latter, we first define the empirical loss function of $\bar{f}$ as

$$\bar{L}(\lambda|\mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N} w_i\{\bar{f}(x_i) - f_0(x_i)\}^2, \tag{11}$$

where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ denotes the collection of all covariates and $w_i \geq 0$ are the associated weights with observation i such that $\sum_{i=1}^N w_i = N$. Similarly, the

corresponding conditional risk $\bar{R}(\lambda|\mathbf{X}) = \mathbb{E}_\varepsilon\{\bar{L}(\lambda|\mathbf{X})\}$ has the following form

$$\bar{R}(\lambda|\mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N} w_i \left[\frac{1}{m}\sum_{k=1}^{m}\left\{\mathbb{E}_\varepsilon \widehat{f}_k(x_i) - f_0(x_i)\right\}\right]^2 + \frac{1}{m^2 N}\sum_{k=1}^{m}\sum_{i=1}^{N} w_i \mathrm{Var}_\varepsilon\left\{\widehat{f}_k(x_i)\right\}. \quad (12)$$

The form of (10) illustrates that, roughly speaking, a "locally optimal" choice of λ (that minimizes (8)) tries to strike a good balance of variance and bias for each sub-estimate $\widehat{f}_k$. On the contrary, a "globally optimal" λ, which is defined to minimize (11), puts much less emphasis on the variance of $\widehat{f}_k$ (by a factor of $1/m$) than on the bias of $\widehat{f}_k$; see (12). Consequently, to obtain a "globally optimal" $\bar{f}$, one needs to intentionally choose a "smaller" λ such that each individual function estimator $\widehat{f}_k$ overfits data set S_k, which leads to reduced bias $\mathbb{E}_\varepsilon \widehat{f}_k(x_i) - f_0(x_i)$ and inflated variance $\mathrm{Var}_\varepsilon\left\{\widehat{f}_k(x_i)\right\}$. Then by taking $\bar{f} = \frac{1}{m}\sum_{j=1}^{m}\widehat{f}_j$, the variance of $\bar{f}$ can be effectively reduced by a factor of $1/m$ while keeping its bias at the same level as those of individual $\widehat{f}_j$'s. The above risk analysis confirms the heuristics in Zhang et al. (2015).

3.3 Distributed Generalized Cross-Validation

The discussions in Section 3.2 motivate the main result of this paper: distributed GCV score, denoted by dGCV. This data-driven tool in selecting λ is computationally efficient for massive data as analyzed in Section 3.4.

Using the solution (6), it is straightforward to show that the predicted values of all data points $\mathbf{y}_l$ in the subset S_l using $\widehat{f}_k$ take the form $\widehat{\mathbf{y}}_{kl} = \mathbf{A}_{kl}\mathbf{y}_k$, where $\mathbf{A}_{kl}(\lambda) = \mathbf{K}_{kl}^T(\mathbf{K}_{kk} + n_k\lambda\mathbf{I}_k)^{-1}$. Define the pooled vector of responses $\mathbf{Y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_m^T)^T$. Then the predicted value of $\mathbf{Y}$ using the averaged estimator $\bar{f}$ is of the form

$$\widehat{\mathbf{Y}} = \left(\frac{1}{m}\sum_{k=1}^{m}\widehat{\mathbf{y}}_{k1}^T, \ldots, \frac{1}{m}\sum_{k=1}^{m}\widehat{\mathbf{y}}_{km}^T\right)^T = \bar{\mathbf{A}}_m(\lambda)\mathbf{Y},$$

where the averaged hat matrix $\bar{\mathbf{A}}_m(\lambda)$ is defined as follows

$$\bar{\mathbf{A}}_m(\lambda) = \frac{1}{m}\begin{pmatrix} \mathbf{A}_{11}(\lambda) & \mathbf{A}_{12}(\lambda) & \cdots & \mathbf{A}_{1m}(\lambda) \\ \mathbf{A}_{21}(\lambda) & \mathbf{A}_{22}(\lambda) & \cdots & \mathbf{A}_{2m}(\lambda) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{m1}(\lambda) & \mathbf{A}_{m2}(\lambda) & \cdots & \mathbf{A}_{mm}(\lambda) \end{pmatrix}. \quad (13)$$

Furthermore, the global conditional risk function (12) can be conveniently re-written as

$$\bar{R}(\lambda|\mathbf{X}) = \frac{1}{N}\boldsymbol{F}^T\{\mathbf{I} - \bar{\mathbf{A}}_m(\lambda)\}^T\mathbf{W}\{\mathbf{I} - \bar{\mathbf{A}}_m(\lambda)\}\boldsymbol{F} + \frac{\sigma^2}{N}\mathrm{tr}\left\{\bar{\mathbf{A}}_m^T(\lambda)\mathbf{W}\bar{\mathbf{A}}_m(\lambda)\right\}, \quad (14)$$

where vector of true values $\boldsymbol{F} = (\mathbf{f}_1^T, \ldots, \mathbf{f}_m^T)^T$ and $\mathbf{W} = \mathrm{diag}\{w_1, \ldots, w_N\}$. Obviously the risk function above cannot be used to select λ in practice since the vector $\boldsymbol{F}$ is

unknown. Following Gu (2013), we can define an unbiased estimator of $\bar{R}(\lambda|\mathbf{X}) + \sigma^2$ as follows

$$\bar{U}(\lambda|\mathbf{X}) = \frac{1}{N}\mathbf{Y}^T\{\mathbf{I} - \bar{\mathbf{A}}_m(\lambda)\}^T\mathbf{W}\{\mathbf{I} - \bar{\mathbf{A}}_m(\lambda)\}\mathbf{Y} + \frac{2\sigma^2}{N}\mathrm{tr}\left\{\bar{\mathbf{A}}_m(\lambda)\mathbf{W}\right\}. \tag{15}$$

It is straightforward to show that $\mathbb{E}_\varepsilon\{\bar{U}(\lambda|\mathbf{X})\} = \bar{R}(\lambda|\mathbf{X}) + \sigma^2$. The above $\bar{U}(\lambda|\mathbf{X})$ can be viewed as an extension of the Mallow's CP (Mallows, 2000) to the divide-and-conquer scenario.

Similar to Gu (2013); Xu and Huang (2012), the Lemma 1 in Section 4 states that under some mild conditions, minimizing $\bar{U}(\lambda|\mathbf{X})$ and $\bar{L}(\lambda|\mathbf{X})$ with respect to λ is asymptotically equivalent. In this sense, the λ chosen by minimizing $\bar{U}(\lambda|\mathbf{X})$ is therefore "globally optimal." However, a major drawback of $\bar{U}(\lambda|\mathbf{X})$ is that it utilizes the knowledge of σ^2, which in practice often needs to be estimated. To overcome this, we propose the following modification of the GCV score

$$\mathrm{dGCV}(\lambda|\mathbf{X}) = \frac{\frac{1}{N}\sum_{i=1}^N w_i\left\{y_i - \bar{f}(x_i)\right\}^2}{\left[1 - \frac{1}{Nm}\sum_{k=1}^m \mathrm{tr}\{\mathbf{A}_{kk}(\lambda)\mathbf{W}_k\}\right]^2}, \tag{16}$$

where $\mathbf{W}_k = \mathrm{diag}\{w_i, i \in S_k\}$. Intuitively, consider $\tilde{\sigma}^2 = N^{-1}\sum_{i=1}^N w_i\left\{y_i - \bar{f}(x_i)\right\}^2$ as an estimator of σ^2 and use the fact that $(1-x)^{-2} \approx 1+2x$ as $x \to 0$, the $\bar{U}(\lambda|\mathbf{X})$ defined in (15) essentially can be viewed as the first order Taylor expansion of the $\mathrm{dGCV}(\lambda|\mathbf{X})$. However, in the definition of $\mathrm{dGCV}(\lambda|\mathbf{X})$, it does not require any information of σ^2. Note that dGCV incorporates information across all sub-samples, which explains its superior empirical performance. In fact, Theorem 1 in Section 4 shows that under some conditions, minimizing $\mathrm{dGCV}(\lambda|\mathbf{X})$ and the "golden criterion" $\bar{L}(\lambda|\mathbf{X})$ with respect to λ are also asymptotically equivalent.

3.4 Computational Complexity of dGCV

The computation of $\mathrm{dGCV}(\lambda|\mathbf{X})$ in (16) for a given λ consists of two parts: the first part involves computing the trace of individual hat matrices, $\mathrm{tr}\{\mathbf{A}_{kk}(\lambda)\mathbf{W}_k\}$, $k = 1, \ldots, m$, which requires $O(N^3/m^2)$ floating operations and a memory usage of $O(N^2/m^2)$; the second part is to evaluate the predicted value of $\bar{f}(x_i)$ for which $w_i \neq 0$, which costs $O(NN_w)$ floating operations and a memory usage of $O(N)$, where N_w denotes the number of nonzero w_i's. Hence, the total computation cost of $\mathrm{dGCV}(\lambda|\mathbf{X})$ is of the order $O(N^3/m^2 + NN_w)$. In cases when $m/\sqrt{N} = O(1)$, one can simply use $w_1 = \cdots = w_N = 1$, which results in the computational cost of the order $O(N^3/m^2)$ for one evaluation of $\mathrm{dGCV}(\lambda|\mathbf{X})$. This is the same as that of the divide-and-conquer algorithm proposed in Zhang et al. (2015).

In some applications where m is much larger than $\sqrt{N}$, the computational cost of $\mathrm{dGCV}(\lambda|\mathbf{X})$ becomes $O(NN_w)$. In this case, we may want to only choose m^* out of m

sub-data sets for saving computational costs. To achieve that, we need to choose weights w_i's properly. For example, we can set $w_i = N/(\sum_{k=1}^{m^*} n_k)$ if $i \in \cup_{k=1}^{m^*} S_k$ and $w_i = 0$ otherwise. Under this setting, the dGCV($\lambda|\mathbf{X}$) in (16) becomes

$$\text{dGCV}^*(\lambda|\mathbf{X}) = \frac{\frac{1}{N_{m^*}} \sum_{i \in \cup_{k=1}^{m^*} S_k} \left\{ y_i - \bar{f}(x_i) \right\}^2}{\left[1 - \frac{1}{m N_{m^*}} \sum_{k=1}^{m^*} \text{tr}\{\mathbf{A}_{kk}(\lambda)\} \right]^2}, \tag{17}$$

where $N_{m^*} = n_1 + \cdots + n_{m^*}$. Using (17) instead of (16), we only need to evaluate $\bar{f}(x_i)$ for x_i's in m^* subsets and the computation time is reduced to $O(N^2 m^*/m + N^3/m^2)$. We applied (17) to the Million Song Data set considered in Section 6, which yields good results in both prediction and computation time.

Optimization of dGCV($\lambda|\mathbf{X}$) or dGCV*($\lambda|\mathbf{X}$) can be carried out using a simple one-dimensional grid search. Since the first and second derivatives of dGCV($\lambda|\mathbf{X}$) or dGCV*($\lambda|\mathbf{X}$) can be easily computed using similar arguments in Wood (2004); Xu and Huang (2012), it can also be optimized using the Newton-Raphson algorithm with the same computational costs.

3.5 The Newton-Raphson Implementation

In some applications, not only the penalty parameter λ in (2) needs to be carefully selected, it is also important to choose other tuning parameters in the kernel function. For example, the bandwidth parameter ϕ in the Gaussian kernel $K(\boldsymbol{x}, \boldsymbol{z}) = \exp(-\|\boldsymbol{x} - \boldsymbol{z}\|_2^2/\phi)$ also plays an important role in the performance of the KRR, as we will illustrate in the Million Song Dataset in Section 6. In such cases, dGCV can serve as a tool to choose the optimal tuning parameters $\boldsymbol{\theta}$ in the kernel function, as long as conditions C1-C4 in Section 4.1 are satisfied. One remaining practical issue is that when the dimension of $\boldsymbol{\theta}$ is high, the grid search method for the optimal combination of λ and $\boldsymbol{\theta}$ using dGCV is no longer feasible. Therefore, it is necessary to develop more efficient algorithms such as the Newton-Raphson type algorithm.

Following Wood (2004), denote $\eta = \log \lambda$ and $\text{dGCV}(\eta, \boldsymbol{\theta}) = \alpha(\eta, \boldsymbol{\theta})/\gamma(\eta, \boldsymbol{\theta})$, where

$$\begin{aligned} \alpha(\eta, \boldsymbol{\theta}) &= \frac{1}{N} \mathbf{Y}^T \{\mathbf{I} - \bar{\mathbf{A}}_m(\eta, \boldsymbol{\theta})\}^T \mathbf{W} \{\mathbf{I} - \bar{\mathbf{A}}_m(\lambda, \boldsymbol{\theta})\} \mathbf{Y}, \\ \gamma(\eta, \boldsymbol{\theta}) &= \left[1 - \frac{1}{Nm} \sum_{k=1}^{m} \text{tr}\{\mathbf{A}_{kk}(\eta, \boldsymbol{\theta}) \mathbf{W}_k\} \right]^2, \end{aligned}$$

with $\bar{\mathbf{A}}_m(\eta, \boldsymbol{\theta})$ and $\mathbf{A}_{kk}(\eta, \boldsymbol{\theta})$'s defined in (13). Then the first and second partial derivatives of $\log[\text{dGCV}(\eta, \boldsymbol{\theta})]$ can be straightforwardly obtained as

$$\frac{\partial \log[\text{dGCV}(\eta, \boldsymbol{\theta})]}{\partial \vartheta} = \frac{1}{\alpha(\eta, \boldsymbol{\theta})} \frac{\partial \alpha(\eta, \boldsymbol{\theta})}{\partial \vartheta} - \frac{1}{\gamma(\eta, \boldsymbol{\theta})} \frac{\partial \gamma(\eta, \boldsymbol{\theta})}{\partial \vartheta}, \qquad \vartheta = \eta \text{ or } \boldsymbol{\theta}.$$

$$\frac{\partial^2 \log\left[\mathrm{dGCV}(\eta,\boldsymbol{\theta})\right]}{\partial\vartheta\partial\varrho^T} = -\frac{1}{\alpha^2(\eta,\boldsymbol{\theta})}\left[\frac{\partial\alpha(\eta,\boldsymbol{\theta})}{\partial\vartheta}\right]\left[\frac{\partial\alpha(\eta,\boldsymbol{\theta})}{\partial\varrho}\right]^T + \frac{1}{\alpha(\eta,\boldsymbol{\theta})}\frac{\partial^2\alpha(\eta,\boldsymbol{\theta})}{\partial\vartheta\partial\varrho^T}$$
$$+ \frac{1}{\gamma^2(\eta,\boldsymbol{\theta})}\left[\frac{\partial\gamma(\eta,\boldsymbol{\theta})}{\partial\vartheta}\right]\left[\frac{\partial\gamma(\eta,\boldsymbol{\theta})}{\partial\varrho}\right]^T - \frac{1}{\gamma(\eta,\boldsymbol{\theta})}\frac{\partial^2\gamma(\eta,\boldsymbol{\theta})}{\partial\vartheta\partial\varrho^T}, \qquad \vartheta,\varrho = \eta \text{ or } \boldsymbol{\theta}.$$

By definitions of $\alpha(\eta,\boldsymbol{\theta})$ and $\gamma(\eta,\boldsymbol{\theta})$, straightforward matrix calculus yields that it remains to compute partial derivatives of $\mathbf{A}_{kl}(\eta,\boldsymbol{\theta}) = \mathbf{K}_{kl}^T(\boldsymbol{\theta})\left[\mathbf{K}_{kk}(\boldsymbol{\theta}) + n_k e^{\eta}\mathbf{I}_k\right]^{-1}$ with $\mathbf{K}_{kl} = \left[K(\boldsymbol{x}_i,\boldsymbol{x}_j;\boldsymbol{\theta})\right]_{i\in S_k, j\in S_l}$ for $l,k = 1,\ldots,m$. It is straightforward to show that

$$\frac{\partial\mathbf{A}_{kl}(\eta,\boldsymbol{\theta})}{\partial\eta} = -n_k e^{\eta}\mathbf{K}_{kl}^T(\boldsymbol{\theta})\mathbf{K}_{kk}^{\ddagger 2}, \qquad \frac{\partial\mathbf{A}_{kl}(\eta,\boldsymbol{\theta})}{\partial\theta_c} = \frac{\partial\mathbf{K}_{kl}^T(\boldsymbol{\theta})}{\partial\theta_c}\mathbf{K}_{kk}^{\ddagger} - \mathbf{K}_{kl}^T(\boldsymbol{\theta})\mathbf{K}_{kk}^{\ddagger}\frac{\partial\mathbf{K}_{kk}(\boldsymbol{\theta})}{\partial\theta_c}\mathbf{K}_{kk}^{\ddagger},$$
$$\frac{\partial^2\mathbf{A}_{kl}(\eta,\boldsymbol{\theta})}{\partial\eta^2} = -n_k e^{\eta}\mathbf{K}_{kl}^T(\boldsymbol{\theta})\mathbf{K}_{kk}^{\ddagger 2} + 2n_k^2 e^{2\eta}\mathbf{K}_{kl}^T(\boldsymbol{\theta})\mathbf{K}_{kk}^{\ddagger 3},$$
$$\frac{\partial^2\mathbf{A}_{kl}(\eta,\boldsymbol{\theta})}{\partial\eta\partial\theta_c} = -n_k e^{\eta}\left\{\frac{\partial\mathbf{K}_{kl}^T(\boldsymbol{\theta})}{\partial\theta_c} - \mathbf{K}_{kl}^T(\boldsymbol{\theta})\mathbf{K}_{kk}^{\ddagger}\frac{\partial\mathbf{K}_{kk}(\boldsymbol{\theta})}{\partial\theta_c}\right\}\mathbf{K}_{kk}^{\ddagger 2} + n_k e^{\eta}\mathbf{K}_{kl}^T(\boldsymbol{\theta})\mathbf{K}_{kk}^{\ddagger 2}\frac{\partial\mathbf{K}_{kk}(\boldsymbol{\theta})}{\partial\theta_c}\mathbf{K}_{kk}^{\ddagger},$$
$$\begin{aligned}\frac{\partial^2\mathbf{A}_{kl}(\eta,\boldsymbol{\theta})}{\partial\theta_{c_1}\partial\theta_{c_2}} = {} & \frac{\partial^2\mathbf{K}_{kl}^T(\boldsymbol{\theta})}{\partial\theta_{c_1}\partial\theta_{c_2}}\mathbf{K}_{kk}^{\ddagger} - \frac{\partial\mathbf{K}_{kl}^T(\boldsymbol{\theta})}{\partial\theta_{c_1}}\mathbf{K}_{kk}^{\ddagger}\frac{\partial\mathbf{K}_{kk}^T(\boldsymbol{\theta})}{\partial\theta_{c_2}}\mathbf{K}_{kk}^{\ddagger} - \frac{\partial\mathbf{K}_{kl}^T}{\partial\theta_{c_2}}(\boldsymbol{\theta})\mathbf{K}_{kk}^{\ddagger}\frac{\partial\mathbf{K}_{kk}(\boldsymbol{\theta})}{\partial\theta_{c_1}}\mathbf{K}_{kk}^{\ddagger} \\ & + \mathbf{K}_{kl}^T(\boldsymbol{\theta})\mathbf{K}_{kk}^{\ddagger}\frac{\partial\mathbf{K}_{kk}(\boldsymbol{\theta})}{\partial\theta_{c_2}}\mathbf{K}_{kk}^{\ddagger}\frac{\partial\mathbf{K}_{kk}(\boldsymbol{\theta})}{\partial\theta_{c_1}}\mathbf{K}_{kk}^{\ddagger} - \mathbf{K}_{kl}^T(\boldsymbol{\theta})\mathbf{K}_{kk}^{\ddagger}\frac{\partial^2\mathbf{K}_{kk}(\boldsymbol{\theta})}{\partial\theta_{c_1}\partial\theta_{c_2}}\mathbf{K}_{kk}^{\ddagger} \\ & + \mathbf{K}_{kl}^T(\boldsymbol{\theta})\mathbf{K}_{kk}^{\ddagger}\frac{\partial\mathbf{K}_{kk}(\boldsymbol{\theta})}{\partial\theta_{c_1}}\mathbf{K}_{kk}^{\ddagger}\frac{\partial\mathbf{K}_{kk}(\boldsymbol{\theta})}{\partial\theta_{c_2}}\mathbf{K}_{kk}^{\ddagger}, \qquad \text{for } \boldsymbol{\theta} = (\theta_1,\cdots,\theta_D), c, c_1, c_2 = 1,\cdots,D,\end{aligned}$$

where $\mathbf{K}_{kk}^{\ddagger} = \left[\mathbf{K}_{kk}(\boldsymbol{\theta}) + n_k e^{\eta}\mathbf{I}_k\right]^{-1}$, $k,l = 1,\cdots,m$ and all matrix derivatives are taken element-wise.

It is straightforward to show that the computational complexity of first and second derivatives of $\log\left[\mathrm{dGCV}(\eta,\boldsymbol{\theta})\right]$ are the same as that of dGCV, which makes the Newton-Raphson type algorithm feasible. However, it is worth pointing out that $\log\left[\mathrm{dGCV}(\eta,\boldsymbol{\theta})\right]$ is not a convex function of η and $\boldsymbol{\theta}$, hence there is no guarantee that a Newton-Raphson type algorithm will converge to the global minimizer. Numerical suggestions such as those in Wood (2004) may be useful for developing more efficient algorithms, which will be an interesting further research topic.

4 Asymptotic Properties

In this section, we will show that the proposed dGCV criterion in (16) is "globally optimal" under some conditions. We first introduce some notation. Denote $\mathbb{P}_X$, $\mathbb{P}_\varepsilon$, $\mathbb{P}_{\varepsilon,X}$ as the probability measures of covariate X, error process ε and their joint probability measure. Similarly, $\mathbb{E}_\varepsilon$ and Var_ε denote the expectation and variance under the probability measure $\mathbb{P}_\varepsilon$. Let $\lambda_{\max}(\mathbf{A})$ and $\sigma_{\max}(\mathbf{A})$ and $\mathrm{tr}(\mathbf{A})$ be the largest eigenvalue and the largest singular value of the matrix $\mathbf{A}$, respectively. We use $\xrightarrow{\mathbb{P}}$ to denote the convergence in probability measure $\mathbb{P}$ and $O_{\mathbb{P}}(\cdot)$, $o_{\mathbb{P}}(\cdot)$ as defined in the conventional way. For any function $f(x): \mathcal{X} \to \mathbb{R}$, let $\|f\|_{\sup} = \sup_{x\in\mathcal{X}}|f(x)|$ and $\mathbb{P}f = \int_{\mathcal{X}} f(x)\,d\mathbb{P}$.

Finally, let $\mathbb{P}_n$ denote the empirical probability measure based on i.i.d samples of size n from the probability measure $\mathbb{P}$.

4.1 Asymptotic Optimality of dGCV

The following regularity conditions are needed to show the optimality of dGCV.

[C1] $\frac{1}{m}\sum_{l=1}^{m}\lambda_{\max}\left\{(\mathbf{K}_{ll}+n_l\lambda\mathbf{I}_l)^{-2}\left(\frac{1}{m}\sum_{k=1}^{m}\mathbf{K}_{kl}^T\mathbf{K}_{kl}\right)\right\}=O_{\mathbb{P}_X}(1)$;

[C2] $N\bar{R}(\lambda|\mathbf{X})\xrightarrow{\mathbb{P}_X}\infty$ as $N\to\infty$;

[C3] (a) The weights w_i's are nonnegative such that $\sum_{i=1}^{N}w_i=N$ and that $\max_{1\le i\le N}w_i\le W$ for some constant $W>0$; (b) $\frac{1}{Nm}\sum_{k=1}^{m}\mathrm{tr}\{\mathbf{A}_{kk}(\lambda)\}=o_{\mathbb{P}_X}(1)$ as $N\to\infty$.

[C4] $\frac{[N^{-1}\mathrm{tr}\{\bar{\mathbf{A}}_m(\lambda)\mathbf{W}\}]^2}{\left[N^{-1}\mathrm{tr}\{\bar{\mathbf{A}}_m^T(\lambda)\mathbf{W}\bar{\mathbf{A}}_m(\lambda)\}\right]}=o_{\mathbb{P}_X}(1)$ as $N\to\infty$.

Intuitively, condition C1 requires that some similarities among sub-data sets. If all $\mathbf{K}_{kl}$'s are similar to $\mathbf{K}_{ll}$, we can expect $\lambda_{\max}\left\{(\mathbf{K}_{ll}+n_l\lambda\mathbf{I}_l)^{-2}\left(\frac{1}{m}\sum_{k=1}^{m}\mathbf{K}_{kl}^T\mathbf{K}_{kl}\right)\right\}\le 1$, in which case C1 holds. In Section 4.2, we shall show that one sufficient condition for C1 to hold is to ensure that the "maximal marginal degrees of freedom" (Bach, 2013) d_λ defined in (20) is sufficiently small compared to N/m. Condition C2 is a widely used condition to ensure the optimality of the GCV to hold, for example, see Craven and Wahba (1978); Li (1986); Gu and Ma (2005); Xu and Huang (2012). It is a mild condition for nonparametric regression problems, where the parametric rate $O(N^{-1})$ is unattainable for the estimation risk. For example, for kernel ridge regression models with polynomially or exponentially decaying kernel functions, condition C2 holds (Zhang et al., 2015). However, it does raise a flag for the application of the dGCV when a finite rank kernel is used, in which case the optimal rate of $\bar{R}(\lambda|\mathbf{X})$ is of the order $O(N^{-1})$ (Zhang et al., 2015). Nevertheless, without condition C2, it is questionable whether there exists an asymptotically optimal selection procedure for the tuning parameter λ (Li, 1986).

Remark 1. *Condition C3(a) has an important implication for the dGCV*(λ) defined in Section 3.4. When leaving out a portion of data as suggested in Section 3.4, the resulting weights become $w_i=N/(\sum_{k=1}^{m^*}n_k)$ if $i\in\cup_{k=1}^{m^*}S_k$ and $w_i=0$ otherwise. Condition C3(a) requires that the number of data points remained (i.e., $\sum_{k=1}^{m^*}n_k$) must be of the same order as N. Therefore, more data points need to be retained as the sample size N grows. Furthermore, when all sub-datasets under the divide-and-conquer procedure are roughly of the same size, Condition C3(a) essentially requires that $m^*/m=c$ for some absolute constant $0<c\le 1$. From the computational point of view, it is worth to use a $m^*<m$ only when $N>>m^2$. Therefore, a general rule of thumb for the choice of m^* is that it should only be used when $N>>m^2$ and if used it cannot be too small compared to m.*

It turns out that, under conditions C1-C2 and C3(a), $\bar{U}(\lambda|\mathbf{X})$ defined in (15) is "globally optimal."

Lemma 1. *Under Conditions C1–C2 and C3(a), for a fixed λ, we have that*

$$\bar{U}(\lambda|\mathbf{X}) - \bar{L}(\lambda|\mathbf{X}) - \frac{1}{N}\boldsymbol{\varepsilon}^T\mathbf{W}\boldsymbol{\varepsilon} = o_{\mathbb{P}_{\varepsilon,X}}\{\bar{L}(\lambda|\mathbf{X})\}. \tag{18}$$

The proof is given in the Appendix.

Lemma 1 states that when σ^2 is known, minimizing $\bar{U}(\lambda|\mathbf{X})$ with respect to λ is asymptotically equivalent to minimizing the empirical true loss function $\bar{L}(\lambda|\mathbf{X})$. However, it is rarely the case that one has complete knowledge of σ^2. In this sense, the proposed dGCV is more practical and it can be shown to be "globally optimal" as well, under some additional conditions.

Theorem 1. *Under Conditions C1–C4, for a fixed λ, we have that*

$$\mathrm{dGCV}(\lambda|\mathbf{X}) - \bar{L}(\lambda|\mathbf{X}) - \frac{1}{N}\boldsymbol{\varepsilon}^T\mathbf{W}\boldsymbol{\varepsilon} = o_{\mathbb{P}_{\varepsilon,X}}\{(\bar{L}(\lambda|\mathbf{x})\}. \tag{19}$$

The proof is given in the Appendix.

Similar to Lemma 1, Theorem 1 shows that minimizing $\mathrm{dGCV}(\lambda|\mathbf{X})$ amounts to minimizing the true conditional loss function $\bar{L}(\lambda|\mathbf{X})$, although additional conditions C3(b)-C4 are needed. Condition C3(b) is rather mild in that it essentially requires that the effective degrees of freedom to be negligible compared to the sample size, which is typically true for non-parametric function estimators in most settings of interest. In addition, C3(b) becomes trivial when $m \to \infty$ because by definition we have that $\mathrm{tr}\{\mathbf{A}_{kk}(\lambda)\} \leq n_k$, $k = 1, \ldots, m$. When the entire data set is used at once ($m = 1$), condition C4 reduces to the well known condition $[N^{-1}\mathrm{tr}\{\mathbf{A}(\lambda)\}]^2/[N^{-1}\mathrm{tr}\{\mathbf{A}^2(\lambda)\}] = o(1)$ in the literature (Craven and Wahba, 1978; Li, 1986; Gu and Ma, 2005; Xu and Huang, 2012). For example, for smoothing splines, we typically have $\mathrm{tr}\{\mathbf{A}(\lambda)\} = O(\lambda^{-1/s})$ and $\mathrm{tr}\{\mathbf{A}^2(\lambda)\} \asymp O(\lambda^{-1/s})$ for some $s > 1$. Then as long as $\lambda^{-1/s}/N \to 0$, which covers the most region of practical interest for λ, we have that $[N^{-1}\mathrm{tr}\{\mathbf{A}(\lambda)\}]^2/[N^{-1}\mathrm{tr}\{\mathbf{A}^2(\lambda)\}] \to 0$ as $N \to \infty$. Condition C4 can be viewed as an extension of this commonly used condition to the divide-and-conquer regime.

4.2 Low-level Sufficient Conditions for C1 and C4

In this subsection, for simplicity, we only consider uniform weights with $w_1 = \cdots = w_N = 1$ and equal sample sizes $n_1 = \cdots = n_m = n$ in this subsection. We first establish a low-level sufficient condition for C1. Following Bach (2013), define the "maximal marginal degrees of freedom" as

$$d_\lambda = N\|\mathrm{diag}\{\mathbf{K}(\mathbf{K} + N\lambda\mathbf{I}_N)^{-1}\}\|_\infty, \tag{20}$$

where $\|\cdot\|_\infty$ stands for the matrix infinity norm. Note that $\mathbf{A}(\lambda) = \mathbf{K}(\mathbf{K} + N\lambda\mathbf{I}_N)^{-1}\}$ is the hat matrix (13) with $m = 1$ and $\mathrm{df}_\lambda = \mathrm{tr}\left[\mathbf{A}(\lambda)\right] = \|\mathrm{diag}\{\mathbf{K}(\mathbf{K}+N\lambda\mathbf{I}_N)^{-1}\}\|_1$ defines the "effective degrees of freedom" (Gu, 2013) for the KRR using the entire dataset at once. In this sense, the "maximal marginal degrees of freedom" d_λ provides an upper bound for the "effective degree of freedom" df_λ due to the inequality $\mathrm{df}_\lambda \leq d_\lambda$, and hence gives another measure for the model complexity.

[C1'] Let $r = \mathrm{rank}(\mathbf{K})$ and d_λ be the "maximal marginal degrees of freedom" defined in (20), we assume that

$$\frac{m d_\lambda \left(\log r + \log m\right)}{N} = o_{\mathbb{P}_X}(1), \tag{21}$$

as $N \to \infty$ for either a finite m or $m \to \infty$.

Condition C1' ensures that the number of partitions m cannot be too large compared to the total sample size N, depending on the magnitude of d_λ, which is consistent with findings in the literature (Zhang et al., 2015; Shang and Cheng, 2017). With a large m, condition C1' maybe violated if there is a significant number of outliers, leading to a potentially large d_λ.

Lemma 2. *Condition C1' is sufficient for condition C1.*

The proof is given in the Appendix.

Next we proceed to derive sufficient conditions for condition C4. When the entire data set is used at once ($m = 1$) and conditional on observed covariate $\mathbf{X}$, condition C4 reduces to the well known condition $[N^{-1}\mathrm{tr}\{\mathbf{A}(\lambda)\}]^2/[N^{-1}\mathrm{tr}\{\mathbf{A}^2(\lambda)\}] = o(1)$ in the literature (Craven and Wahba, 1978; Li, 1986; Gu and Ma, 2005; Xu and Huang, 2012). For example, for smoothing splines, we typically have $\mathrm{tr}\{\mathbf{A}(\lambda)\} = O(\lambda^{-1/s})$ and $\mathrm{tr}\{\mathbf{A}^2(\lambda)\} \asymp O(\lambda^{-1/s})$ for some $s > 1$. In this case, as long as $\lambda^{-1/s}/N \to 0$, which covers the most region of practical interest for λ, we have that $[N^{-1}\mathrm{tr}\{\mathbf{A}(\lambda)\}]^2/[N^{-1}\mathrm{tr}\{\mathbf{A}^2(\lambda)\}] \to 0$ as $N \to \infty$. Condition C4 can be viewed as an extension of this commonly used condition to the divide-and-conquer regime, whose justification, however, is much less straightforward.

We first provide some heuristic insights behind our proof. Define

$$Q(\lambda|\mathbf{X}) = \int_{\mathcal{X}} \mathrm{Var}_\varepsilon\{\bar{f}(x)\}^2 \, d\mathbb{P}_X(x) = \frac{1}{m^2}\sum_{k=1}^{m}\int_{\mathcal{X}} \mathrm{Var}_\varepsilon\{\widehat{f}_k(x)\} \, d\mathbb{P}_X(x). \tag{22}$$

Let $\mathbb{P}_{X,N}$ be the empirical measure based on sample $\{X_1, \ldots, X_N\}$, and $\mathbb{P}_{X,n_k}$ be the empirical measure based on the k-th sub-sample $\{X_i\}_{i\in S_k}$. It is straightforward to show that

$$Q_1(\lambda|\mathbf{X}) = \sigma^2\frac{\mathrm{tr}\{\bar{\mathbf{A}}_m^T(\lambda)\bar{\mathbf{A}}_m(\lambda)\}}{N} = \int_{\mathcal{X}} \mathrm{Var}_\varepsilon\left\{\bar{f}(x)\right\}^2 \, d\mathbb{P}_{X,N}(x), \tag{23}$$

$$Q_2(\lambda|\mathbf{X}) = \sigma^2 \frac{1}{Nm} \sum_{k=1}^{m} \mathrm{tr}\{\mathbf{A}_{kk}^2(\lambda)\} = \frac{1}{m^2} \sum_{k=1}^{m} \int_{\mathcal{X}} \mathrm{Var}_\varepsilon\{\widehat{f}_k(x)\}\, d\mathbb{P}_{X,n_k}(x). \tag{24}$$

Intuitively, $Q_1(\lambda|\mathbf{X})$ and $Q_2(\lambda|\mathbf{X})$ are two empirical versions of $Q(\lambda|\mathbf{X})$ and should be close to each other. The formal proof utilizes the uniform ratio limit theorems for empirical processes (Pollard, 1995) to show $Q_1(\lambda|\mathbf{X})/Q(\lambda|\mathbf{X}) = 1 + o_{\mathbb{P}_X}(1)$ and $Q_2(\lambda|\mathbf{X})/Q(\lambda|\mathbf{X}) = 1 + o_{\mathbb{P}_X}(1)$, then with the help of condition C4'(a), we can show condition C4 holds.

Let $\mathcal{N}(\epsilon, \|\cdot\|_{\mathbb{P}_{X,n}}, \mathcal{F})$ be the ϵ-covering number (Pollard, 1986) of a function class $\mathcal{F}$ with the empirical norm $\|f\|_{\mathbb{P}_{X,n}} = \sqrt{n^{-1}\sum_{i=1}^{n} f^2(X_i)}$. Following conditions are sufficient to ensure condition C4.

[C4'](a) $\frac{1}{m}\sum_{k=1}^{m} \left[\frac{1}{N}\mathrm{tr}\{\mathbf{A}_{kk}(\lambda)\}\right]^2 / \left[\frac{1}{N}\mathrm{tr}\{\mathbf{A}_{kk}^2(\lambda)\}\right] = o_{\mathbb{P}_X}(1)$;

[C4'](b) There exists a positive sequence $\{V_n\}$ such that as $V_n \to 0$, it holds that $V_n \left[\frac{1}{m}\sum_{k=1}^{m}\int_{\mathcal{X}} \mathrm{Var}_\varepsilon\{\widehat{f}_k(x)\}\, d\mathbb{P}_X(x)\right]^{-1} = O_{\mathbb{P}_X}(1)$, $\max_{1\le k\le m} \|\mathrm{Var}_\varepsilon\{\widehat{f}_k(x)\}\|_{\sup} = O_{\mathbb{P}_X}(V_n)$ and $nV_n \to \infty$ as $n \to \infty$;

[C4'](c) There exists a sequence $\{H_n\}$ such that $H_n \left[\frac{n}{m}\sum_{k=1}^{m}\int_{\mathcal{X}} \mathrm{Var}_\varepsilon\{\widehat{f}_k(x)\}\, d\mathbb{P}_X(x)\right]^{-1} = O_{\mathbb{P}_X}(1)$, $\max_{1\le k\le m}[\int_{\mathcal{X}} \mathrm{Var}_\varepsilon\{\widehat{f}_k'(x)\}\, d\mathbb{P}_X(x) / \int_{\mathcal{X}} \mathrm{Var}_\varepsilon\{\widehat{f}_k(x)\}\, d\mathbb{P}_X(x)] = O_{\mathbb{P}_X}(H_n^2)$, and $nH_nV_n - (\log m)^2 \to \infty$ as $n \to \infty$. Here, $\widehat{f}_k'(x)$ denotes the derivative of $\widehat{f}_k(x)$;

[C4'](d) For the function class $\mathcal{F}_0 = \{f : \|f\|_{\sup} \le 1, J_1(f) = \int_{\mathcal{X}} \{f'(x)\}^2\, d\mathbb{P}_X(x) \le 1\}$, we have that $\mathcal{N}(\epsilon, \|\cdot\|_{\mathbb{P}_{X,n}}, \mathcal{F}_0) \le \exp(C_0/\epsilon)$ for some constant $C_0 > 0$ with probability approaching one as $n \to \infty$.

Lemma 3. *For a tuning parameter λ satisfying conditions C4'(a)-(d), one has that*

$$\left\{\frac{1}{N} tr(\bar{\mathbf{A}}_m)\right\}^2 / \left\{\frac{1}{N} tr(\bar{\mathbf{A}}_m^T \bar{\mathbf{A}}_m)\right\} = o_{\mathbb{P}_X}(1).$$

The proof is given in the Appendix.

Condition C4'(a) is a mild condition as we have discussed at the beginning of this subsection. Condition C4'(b) essentially states that the supremum norm and the L_1 norm of the variance function $\mathrm{Var}_\varepsilon\{\widehat{f}_k(x)\}$ are of the same order, which is reasonable when all $\mathrm{Var}_\varepsilon\{\widehat{f}_k(x)\}$'s similarly well-behaved within the support of covariate X. In addition, we should restrict our attention to the range of λ such that $n\mathrm{Var}_\varepsilon\{\widehat{f}_k(x)\} \to \infty$, $k = 1, \ldots, m$. Recall the discussion in subsection 3.2, the optimal $\bar{f}$ can only be obtained when the risk (10) is dominated by the variance term $\mathrm{Var}_\varepsilon\{\widehat{f}_k(x)\}$ for each individual $\widehat{f}_k(x)$. Hence, letting $nV_n \to \infty$ is reasonable based on the condition C2. Condition C4'(c) essentially asserts that H_n and nV_n are of the same order. For the smoothing spline case, the derivative $\widehat{f}_k'$ is typically more variable than $\widehat{f}_k$ such that one

can expect $H_n \to \infty$. For example, Rice and Rosenblatt (1983) gives the exact rates of convergence for cubic smoothing spline, that is $\int_{\mathcal{X}} \mathrm{Var}_\varepsilon\{\widehat{f}_k(x)\}\, d\mathbb{P}_X(x) \asymp n^{-1}\lambda^{-1/4}$, $\int_{\mathcal{X}} \mathrm{Var}_\varepsilon\{\widehat{f}_k'(x)\}\, d\mathbb{P}_X(x) \asymp n^{-1}\lambda^{-3/4}$. In this case, we have that $H_n \asymp \lambda^{-1/4}$ and $nV_n \asymp \lambda^{-1/4}$. A thorough theoretical investigation of H_n and V_n is difficult in general, though our simulation study (unreported) suggests condition C4'(c) to be reasonable for many reproducing kernels.

Finally, condition C4'(d) holds when the empirical measure $\mathbb{P}_{X,n}$ is replaced by $\mathbb{P}_X$, see, e.g., van de Geer and van de Geer (2000). One can generally expect it to hold when the sample size n is large. The upper bound of the random covering number $\mathcal{N}(\epsilon, \|\cdot\|_{\mathbb{P}_{X,n}}, \mathcal{F}_0)$ determines the rate of convergence of the empirical processes $Q_1(\lambda|\mathbf{X})$ and $Q_2(\lambda|\mathbf{X})$ to $Q(\lambda|\mathbf{X})$. And it can be relaxed similarly as given in Theorem 2.1 of Pollard (1986).

Remark 2. *One benefit of using high level conditions such as C1, C2 and C4 is that they do not involve the response variable and can be computed efficiently using sample data. To deal with the randomness in covariate X, one can bootstrap/resample/subsample from the observed data, which is especially suitable when the sample size under consideration is extremely large. Through this resampling strategy, one can empirically verify C1, C2 and C4, although rigorous justification of such strategy has not been established and will be an interesting topic for future research.*

5 Simulation studies

In this section, we conduct simulation studies to illustrate the effectiveness of dGCV(λ) in choosing the optimal λ for the d-KRR. The data were simulated from the model

$$y = 2.4 \times \mathrm{beta}(x, 30, 17) + 1.6 \times \mathrm{beta}(x, 3, 11) + \varepsilon, \quad x \in [0, 1], \tag{25}$$

where $\mathrm{beta}(x, a, b)$ is the density function of the Beta(a, b) distribution and $\varepsilon \sim N(0, 3^2)$. The covariate x_i's were independently generated from the uniform distribution over the interval $[0, 1]$. For each simulation run, we first generated a data set of the size $N = mn$ and then randomly partition the data sets into m sub-data sets of equal sizes. The divide-and-conquer estimator $\bar{f}$ was obtained as given in (3).

Let $f^{(\nu)}(\cdot)$ be the νth derivative of a smooth function $f(\cdot)$. The true function in model (25) belongs to the Sobolev Hilbert space of νth order differentiable functions on $[0, 1]$ satisfying the periodic boundary conditions $f^{(\nu)}(0) = f^{(\nu)}(1)$ for $\nu = 1, \cdots, 10$, denoted as $\mathcal{W}_\nu(per)$ (Wahba, 1990). If $\mathcal{W}_\nu(per)$ is endowed with the norm $\|f\|^2_{\mathcal{W}_\nu} = \left\{\int_0^1 f(x)\, dx\right\}^2 + \int_0^1 \{f^{(\nu)}(x)\}^2\, dx$, then it has a reproducing kernel

$$K(x, z) = \frac{(-1)^{\nu-1}}{(2\nu)!} B_{2\nu}([x - z]), \quad x, z \in [0, 1], \tag{26}$$

where $B_{2\nu}(\cdot)$ is the 2νth Bernoulli polynomials (Abramowitz et al., 1972) and $[x]$ is the fractional part of x. In all simulation runs, the tuning parameter λ was selected by a grid search for $\log(\lambda)$ over 30 equally-spaced grid points over the interval $[-10\nu, -5\nu]$. Three approaches were used for the selection of λ: (i) the distributed GCV (dGCV) approach proposed in (16); (ii) (ii) the naive GCV (nGCV) approach where a $\widehat{\lambda}_k$ is selected for each individual $\widehat{f}_k$ by minimizing the sub-GCV score $\mathrm{GCV}_k(\lambda)$ defined in (9) for $k = 1, \cdots, m$ and then the final estimator is obtained by averaging all $\widehat{f}_k$'s; and (iii) the true empirical loss function (TrueLoss) $\bar{L}(\lambda|\mathbf{X})$ defined in (11). The last approach is not practically feasible since it requires the knowledge of the truth f_0. It merely serves as the "golden criterion" to show the effectiveness of other two approaches. For all approaches, we set the weights $w_i = 1$ for all $i = 1, \ldots, N$ and used $\nu = 2$ for the kernel (26) unless otherwise stated.

5.1 Performances with Moderate Sample Sizes

In this subsection, we evaluated performances of the proposed approach with moderate sample sizes $N = 2^i$, $i = 8, 9, 10, 11, 12$. In this setting, it is still possible to obtain the KRR estimator with the entire data set, i.e., $m = 1$, and enables us to evaluate potential loss using the divided-and-conquer approach as opposed to using all data at once.

5.1.1 Computational Complexity and Estimation Accuracies

We first simulate data from model (25) for various sample sizes $N = 2^i$, $i = 8, 9, 10, 11, 12$ and fit the data with divide-and-conquer regression with $m = 1, 2, 4, 8, 16, 32$. Summary statistics based on 100 simulation runs were illustrated in Figure 1(a)-(f). Figure 1(a) illustrates the computational complexity of one evaluation of $\mathrm{dGCV}(\lambda)$. All simulation runs were carried out in the software R (R Core Team, 2018) on a cluster of 100 Linux machines with a total of 100 CPU cores, with each core running at approximately 2 GFLOPS. We can clearly see that by using the divide-and-conquer strategy, the computational time of the dGCV can be greatly reduced compared to the case when all data were used at once (i.e., $m = 1$).

In Figure 1(b)-(c), we give some comparisons of the dGCV method and the nGCV method. Figure 1(b) shows the scatter plot of true empirical losses, as defined in (11), of the function estimators obtained by minimizing $\mathrm{dGCV}(\lambda)$ versus minimizing the unattainable "golden criterion" (11) over 100 simulation runs. As we can see, majority of points are concentrated around the 45^o straight line, which supports our theoretical findings in Theorem 1. On the contrary, Figure 1(c) shows that true empirical losses of the function estimator based on the nGCV approach are generally larger than the minimum possible true losses, indicating that such function estimators are indeed only "locally" optimal but not "globally optimal."

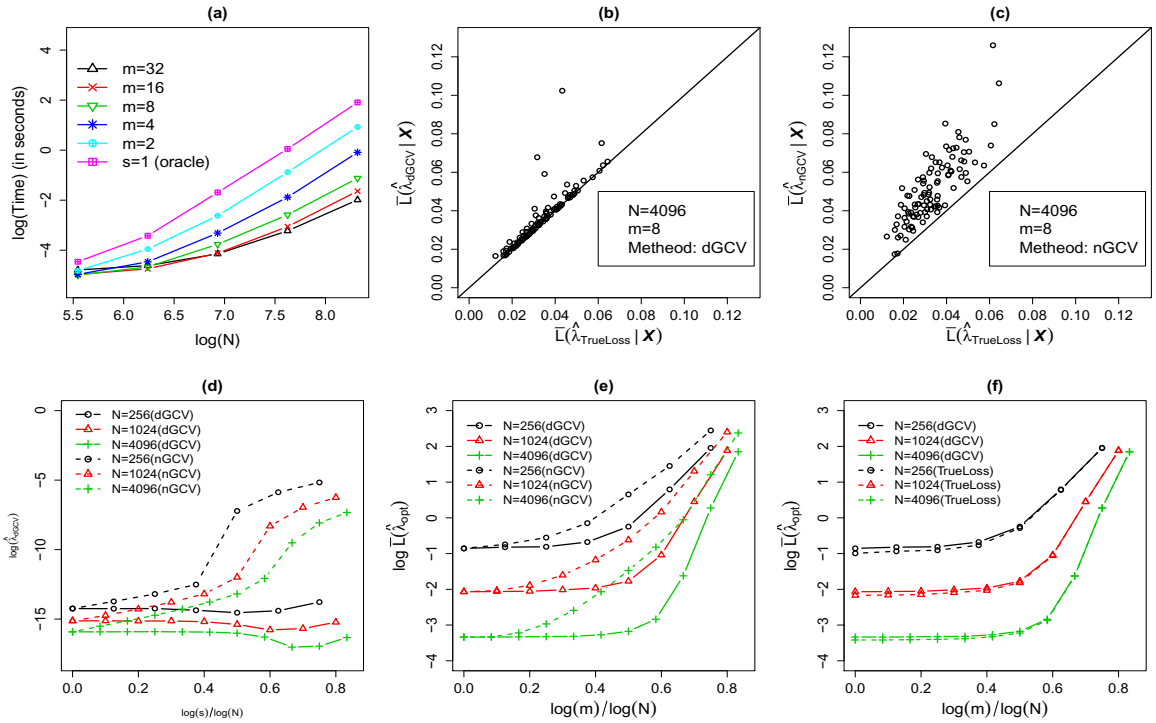

Figure 1: (a) the logarithm of computational time (in seconds) v.s. $\log(N)$; (b)-(c): scatter plots of true empirical losses of function estimators; (d) the logarithm of averages of selected λ v.s. $\log(m)/\log(N)$; (e)-(f): the logarithm of averaged true empirical losses v.s. $\log(m)/\log(N)$. Note that in (d)-(f), $\hat{\lambda}_{\text{opt}}$ in the y-axis denotes one of $\hat{\lambda}_{\text{dGCV}}$, $\hat{\lambda}_{\text{nGCV}}$ and $\hat{\lambda}_{\text{TrueLoss}}$ for each curve.

In Figure 1(d)-(f), we used $N = 2^i$ and $m = 2^j$ for $j = 0, 1, \ldots, i-2$ and $i = 8, 10, 12$ so that there were at least four data points in each sub-data set. To better understand the differences between the dGCV and the nGCV approaches, Figure 1(d) shows how the logarithm of the averages of selected tuning parameters (over 100 simulation runs), denoted as $\log(\widehat{\lambda}_{opt})$, for each method changes as m increases. As we can see, when $m = 1$ they are identical. However, as m increases, the λ selected by the nGCV approach consistently increases whereas the λ selected by the dGCV method stays about the same until m gets really large and is always smaller than the λ selected by the nGCV method. This is consistent with findings in Zhang et al. (2015) where they argue that the locally optimal rate of λ for each individual $\widehat{f}_k$ is of the order $O(n^{-4/5})$ with $n = N/m$ whereas the globally optimal rate for λ is of the order $O(N^{-4/5})$.

The y-axis of Figure 1(e)-(f) is the logarithm of estimation errors $\log \overline{L}(\widehat{\lambda}_{opt})$, where $\overline{L}(\widehat{\lambda}_{opt})$ stands for the averaged true conditional loss defined in (11) over 100 simulation runs using different selection approaches for λ. We can see from Figure 1(e)-(f) that as long as m is not too large compare to N, the proposed dGCV(λ) is quite robust in terms of controlling the estimation error as m grows and is almost identical to that of using the true loss function, which is considered as a "golden criterion." This is consistent with our Theorem 1. In contrast, estimation errors of the nGCV approach quickly inflates as

m increases, which is expected according to our discussion in subsection 3.2. Finally, it is interesting to point out that as the λ selected by the dGCV method starts to drop in Figure 1(d), the estimation errors in Figure 1(e)-(f) start to inflate as well.

5.1.2 Is It Worth Minimizing dGCV(λ)?

In this subsection, we investigate the issue that whether the extra computational costs in minimizing dGCV(λ) is worthwhile. The optimal rates of λ for various reproducing kernels have been well established, see, e.g., Zhang et al. (2015). In the case of the reproducing kernel (26) used in this simulation, the optimal rate for λ is of the order $O\left(N^{-\frac{2\nu}{2\nu+1}}\right)$, or in other words, $\lambda_{opt} = CN^{-\frac{2\nu}{2\nu+1}}$ for some constant C. One misconception is that the choice of C does not matter much because asymptotically any value of C leads to the same convergence rate for $\overline{f}$. However, for a given sample size, this is far from being true. To illustrate, we fitted the data generated from model (25) using reproducing kernel (26) with $\nu = 1$ and 2, respectively. Resulting function estimators based on 100 simulation runs with $N = 2^{12} = 4096$ and $m = 4$ were presented in Figure 2 (a)-(b), where it is apparent that by setting $C = 1$, both KRR estimators based on reproducing kernel with $\nu = 1$ or 2 yield much worse estimation accuracies than those of corresponding KRR estimators using λ selected by minimizing the proposed dGCV(λ) criterion. A closer look at the minimization problem (2), or equivalently (5), suggests that the optimal choice of the constant C in λ_{opt} should depend on (a) the magnitude of the kernel function $K(\cdot, \cdot)$; (b) the magnitude of response $\mathbf{Y}$; (c) the sample size N, and therefore can be difficult to obtain in practice. As we have illustrated in Figure 2, for a fixed sample size, a carefully chosen constant C (through dGCV in this case) may have significant impacts on the quality of resulting KRR estimator, for which reason we believe that additional computational costs in minimizing dGCV is indeed worthwhile.

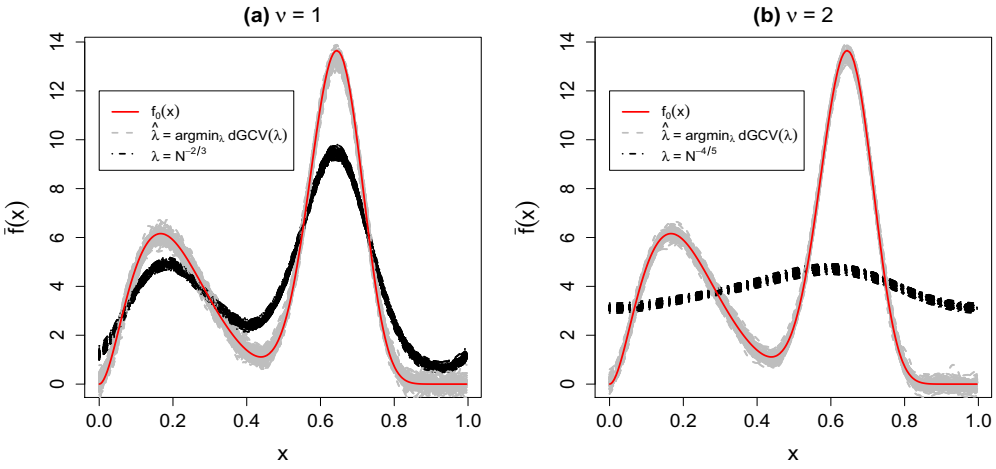

Figure 2: Estimated functions using Divide-and-conquer KRR with a sample size $N = 2^{12}$ and $m = 4$. Kernel defined in (26) was used with (a) $\nu = 1$ and (b) $\nu = 2$.

5.1.3 The Choice of Number of Partitions m

One remaining issue that we have not addressed theoretically is that how many partitions of data (m) should be used in practice for a given sample size N. The general guideline for the choice of m is clear: as long as m is not too large compared to N, the d-KRR estimator can achieve the optimal convergence rate (Zhang et al., 2015; Shang and Cheng, 2017). However, a practical tool to determine whether m is too large is still lacking. In this subsection, we conducted a simulation study to show that the proposed dGCV may serve such a purpose.

By its definition (16), $\mathrm{dGCV}(\lambda)$ can also be viewed as a function of m, denoted as $\mathrm{dGCV}(\lambda, m)$. Then we can define a profiled version of dGCV as follows

$$\mathrm{dGCV}_p(m) = \mathrm{dGCV}(\widehat{\lambda}, m), \tag{27}$$

where $\widehat{\lambda} = \arg\min_{\lambda>0} \mathrm{dGCV}(\lambda, m)$ for a fixed m. We simulated data from model (25) with $N = 2^{12}$ for 100 times and then fitted each data set using d-KRR with $m = 2^j$ for $j = 1, \cdots, 9$. Figure 3(b) presents patterns of 100 centralized version of $\mathrm{dGCV}_p(m)$, defined as $\mathrm{dGCV}_p(m) - \frac{1}{9}\sum_{j=1}^{9} \mathrm{dGCV}_p(j)$, as a function of m. As comparison, Figure 3(a) gives the true empirical loss (11) of each d-KRR estimator using $\widehat{\lambda} = \arg\min_{\lambda>0} \mathrm{dGCV}(\lambda, m)$ for each m, where it appears that as long as $m \leq 2^7$, the estimation accuracy of the fitted function remain roughly the same as using the optimal λ picked by minimizing $\mathrm{dGCV}(\lambda, m)$. This coincides with existing theoretical findings in the literature such as Zhang et al. (2015) and Shang and Cheng (2017). More importantly, the similarity between Figure 3 (a) and (b) suggests that the profiled dGCV score defined in (27) can capture the sudden drop in the trajectory of empirical loss as a function of m and therefore determine which m might be too large. We have tried many other settings and the message remains the same. This implies that, in practical applications, one can start with a relatively large m and gradually decrease m until $\mathrm{dGCV}_p(m)$ defined in (27) stabilizes. Rigorous justifications of such an approach will be an interesting future research topic.

5.1.4 Performances of dGCV on Multivariate Functions

In this subsection, we investigated the impacts of model dimensionality and correlation among predictors on the performance of dGCV. Let $\boldsymbol{x} = (x_1, \cdots, x_p)^T$, the data was simulated from the following model

$$y = f(\boldsymbol{x}) = 20\left(1 - \frac{\|\boldsymbol{x}\|_2}{\sqrt{p}}\right)_+^7 \left(16\frac{\|\boldsymbol{x}\|_2^2}{p} + 7\frac{\|\boldsymbol{x}\|_2}{\sqrt{p}} + 1\right) + \varepsilon, \quad \varepsilon \sim N(0, 3^2), \quad \boldsymbol{x} \in [0,1]^p,$$

where $\|\cdot\|_2$ is the Euclidean norm in $\mathbb{R}^p$, function $(r)_+ = \max(r, 0)$ and x_j's are uniformly distributed between $[0, 1]$ for $j = 1, \cdots, p$. To induce correlations among x_j's, let $x_j = \Phi(z_j)$ where $(z_1, z_2, \cdots, z_p)^T$ was generated from a p-dimensional multivariate normal

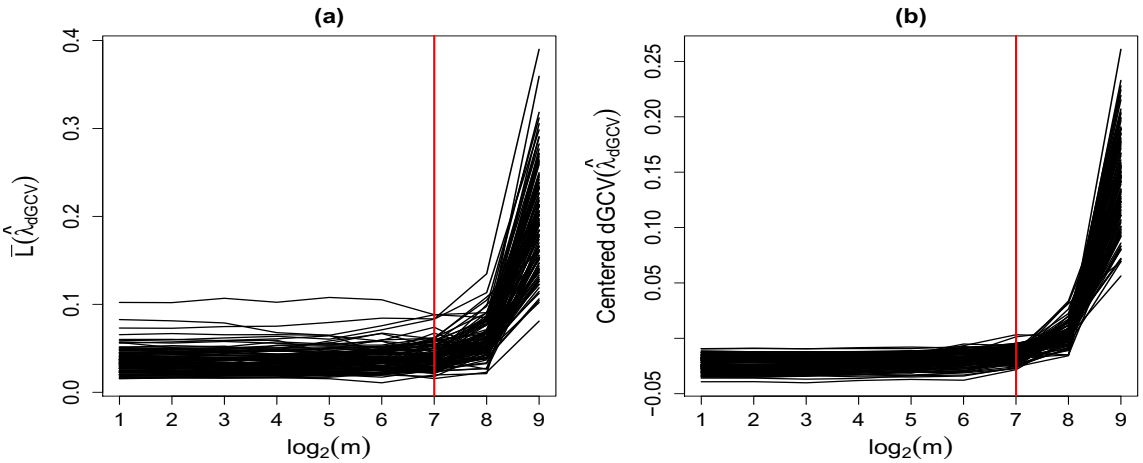

Figure 3: (a) Empirical true loss defined in (11) using λ picked by dGCV for each m; (b) Centered optimal dGCV score for each m; based on 100 simulation runs. ($N = 4096$.)

distribution with mean 0, variance 1 and pairwise correlation coefficient $\rho = 0$ or 0.8. $f(\boldsymbol{x})$ is a variate of Wendland's function (Schaback and Wendland, 2006). For $p \leq 5$, we performed the KRR with the reproducing Hilbert kernel space equipped with the kernel

$$K(\boldsymbol{x}, \boldsymbol{z}) = \left(1 - \frac{\|\boldsymbol{x} - \boldsymbol{z}\|_2}{\sqrt{p}}\right)_+^5 \left(5\frac{\|\boldsymbol{x} - \boldsymbol{z}\|_2^2}{p} + 1\right), \quad \boldsymbol{x}, \boldsymbol{z} \in [0,1]^p,$$

which is a radial basis function with bounded support for $p \leq 5$, see Schaback and Wendland (2006) for more details. The averaged true empirical losses based on 100 simulation runs are summarized in Figure 4. On one hand, when the dimensionality of $\boldsymbol{x}$ increases from $p = 1$ to 5, the averaged empirical losses gradually increase as expected. However, the averaged empirical losses of d-KRR estimators with λ chosen by dGCV is almost indistinguishable from those of corresponding estimators with λ picked by the true empirical loss, regardless of the dimension p. This echoes with our theoretical findings in Theorem 1. On the other hand, as ρ increases from 0 to 0.8, the correlations among x_j's seem to have little impact on the estimation accuracies for the estimated overall mean function $f(\boldsymbol{x})$. In fact, when $\rho = 0.8$, the performance of dGCV is relatively more stable than the case with $\rho = 0$ as the dimension p increases. This can be explained by the fact that $f(\boldsymbol{x})$ only depends on $\|\boldsymbol{x}\|_2$, which is less variable when p increases for the case $\rho = 0.8$. For this reason the estimation of $f(\boldsymbol{x})$ is less affected by the dimensionality when $\rho = 0.8$.

5.2 Performances with a Large Sample Size

In this subsection, we investigated two issues when the sample size N is so large that a single machine can no longer handle at once: (a) whether the computational/estimation performance in Section 5.1.1 still persists; (b) what is the impact of the choice of m^* in (17) on the performance of dGCV*.

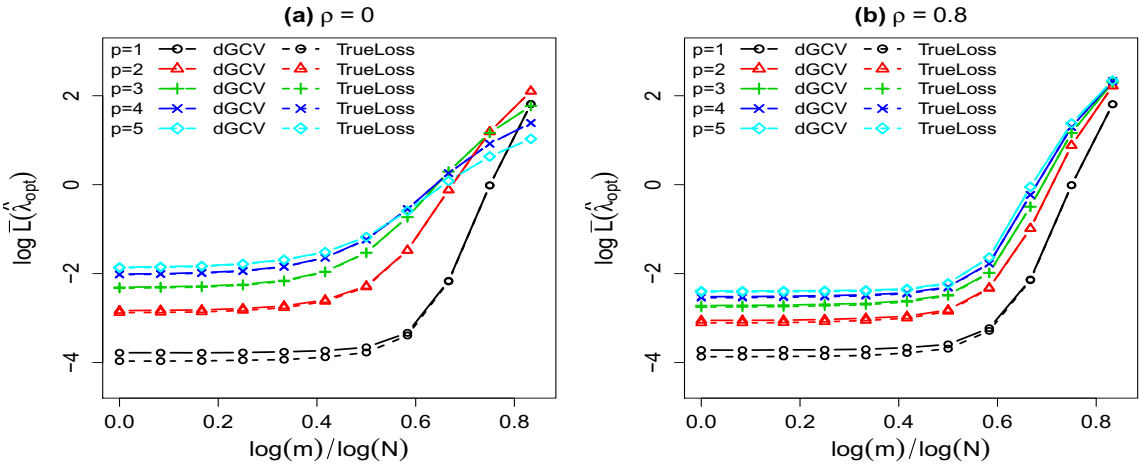

Figure 4: The logarithm of averaged true empirical losses v.s. $\log(m)/\log(N)$ with a sample size $N = 2^{12}$ and (a) $\rho = 0$ (b) $\rho = 0.8$.

5.2.1 Computational Complexity and Estimation Accuracies

To investigate the first issue, we simulated data from model (25) with a sample size $N = 2^{16} = 65,536$ and the d-KRR was carried out using $m = 2^j$ for $j = 5, \cdots, 11$. Summary statistics based on 100 simulation runs are summarized in Figure 5, where the message is consistent with findings presented in Section 5.1.1: at a much smaller computational cost, the d-KRR with a λ chosen by minimizing dGCV is as good as using the λ that minimizes the true empirical loss (11), provided that the m is not too large.

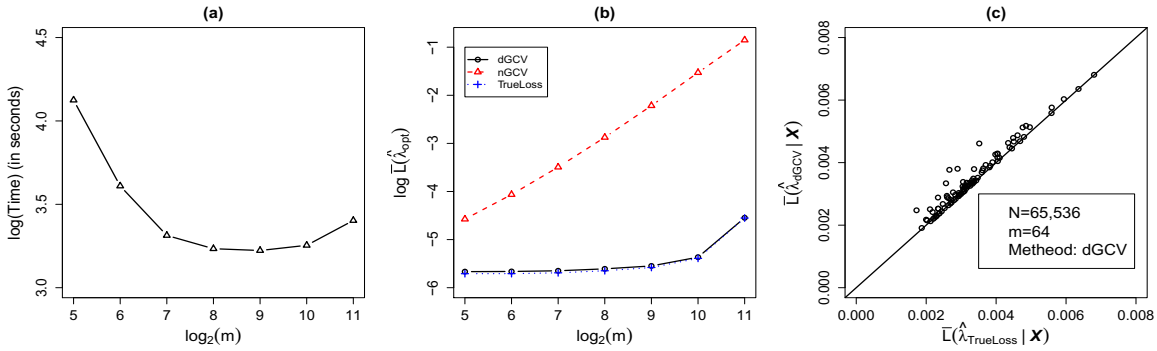

Figure 5: (a) the logarithm of computational time (in seconds) v.s. $\log_2(m)$; (b) the logarithm of averaged true empirical losses v.s. $\log_2(m)$; (c) scatter plots of true empirical losses of function estimators. Note that in (b), $\hat{\lambda}_{\text{opt}}$ in the y-axis denotes one of $\hat{\lambda}_{\text{dGCV}}$, $\hat{\lambda}_{\text{nGCV}}$ and $\hat{\lambda}_{\text{TrueLoss}}$ for each curve.

5.2.2 The Impact of the Choice of m^*

When the sample size N is large or even massive, it is inevitable to use a relative large m, in which case further computational savings can be achieved by choosing a subset of data

for validation as suggested in (17) of Section 3.4. The question remains that how small m^* can be so that Theorem 1 still holds? As we have discussed in Remark 1, a general rule of thumb for the choice of m^* is that it cannot be too small compared to m. To shed some more lights on this issue, for each m, we simulate data from model (25) and then fitted the d-KRR with the λ that minimizes (17) using $m^* = 1, \cdots, m$. Averaged empirical losses based on 100 simulation runs are plotted in Figure 6, where it indicates that if m^* is too small relative to m, the estimation accuracies indeed deteriorate significantly compared to the optimal performance. However, as long as m^* is greater than $0.2m$, the choice of m^* has little impact on the estimation accuracies. Therefore, by setting m^* as a reasonable percentage of m (such as 20% or 30%), one may indeed achieve a large reduction in computational cost without sacrificing too much on estimation accuracies. We want to emphasize again that it is worth to use a $m^* < m$ only when $N >> m^2$. And if used, whenever the computational cost is affordable, a larger m^* is a safer choice to achieve better performances.

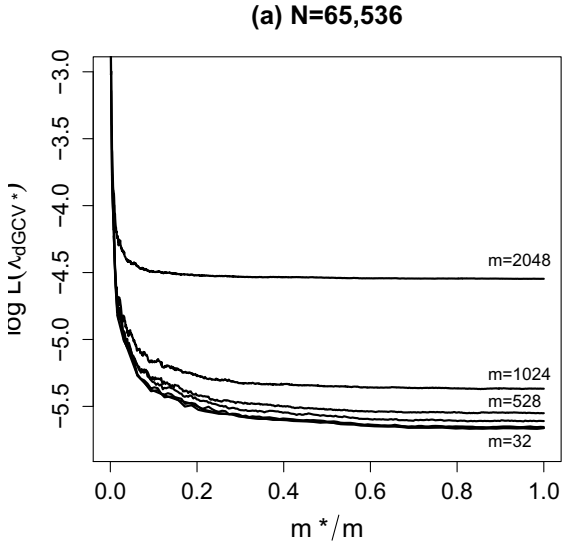

Figure 6: The logarithm of averaged true empirical losses v.s. m^*/m

6 The Million Song Dataset

In this section, we applied the dGCV* tuning method to the Million Song Dataset, which consists of $463,715$ training examples and $51,630$ testing examples. Each observation is a song track released between the year 1922 and 2011. The response variable y_i is the year when the song is released and the covariate x_i is a 90-dimensional vector, consists of timbre information of the song. We refer to Bertin-Mahieux et al. (2011) for more details on this data set. Timbre is the quality of a musical note or sound that distinguishes different types of musical instruments, or voices (Jehan and DesRoches, 2011). The

goal is to use the timbre information of the song to predict the year when the song was released using the KRR. The same dataset has been analyzed by Zhang et al. (2015), but without addressing the issue of selecting an optimal tuning parameter. Our dGCV* method demonstrated significant empirical advantages over theirs.

Following Zhang et al. (2015), the feature vectors were normalized so that they have mean 0 and standard deviation 1 and the Gaussian kernel function $K(x,z) = \exp(-\|x-z\|_2^2/\phi)$ was used for the KRR. Seven partitions $m \in \{32, 38, 48, 64, 96, 128, 256\}$ were used for the d-KRR. Aside from the penalty parameter λ in (2), the bandwidth ϕ is also known to have important impact on the prediction accuracy. To find the best combination of (λ, ϕ) for each partition m, we perform a 2-dimensional search with $\lambda \in \{0.25, 0.5, 0.75, 1.0, 1.25, 1.5\}/N$ and $\phi \in \{2, 3, 4, 5, 6, 7\}$ by minimizing (17) with $m^* = \lceil m/10 \rceil$, where $\lceil a \rceil$ is the smallest integer that is greater than a. See Remark 3 for more details on the choice of m^*. Note that in this case, dGCV$^*(\lambda|\mathbf{X})$ is also a function of ϕ. The experiment was conducted in Matlab using a Windows desktop computer with 32GB of memory and a 2.6Ghz CPU with 4 CPU cores. To illustrate that the computation of the proposed dGCV$^*(\lambda|\mathbf{X})$ can be easily paralleled, Figure 7 gives how averaged computation time changes as the number of CPU cores (in a single machine) increases. The computation time reduces most when the number of CPU cores increases from 1 to 2, and the reductions in computation times slow down as the number of CPU cores continues to increase. Such a trend is probably due to the memory constraints, communication costs and energy consumption limits on the computer and is not uncommon for parallel computing conducted in a single machine. Nevertheless, these computation times are reasonable for a data set with almost half-million observations and can be further reduced if a computing cluster is available.

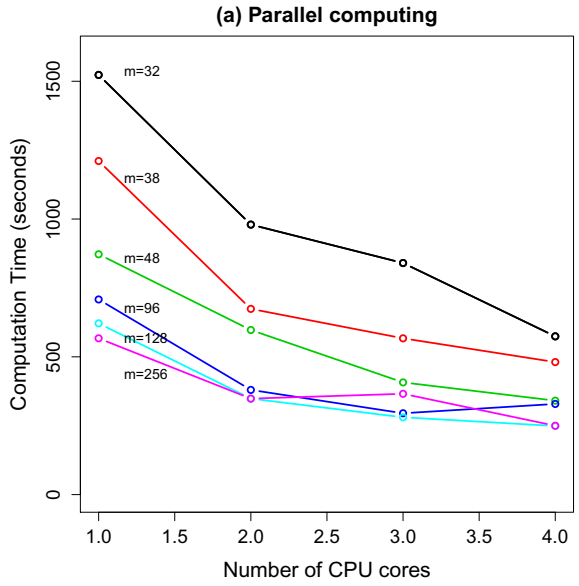

Figure 7: Computing times v.s. Number of CPU cores

The grid search gave the optimal choice of $\lambda = 0.5/N$ and $\phi = 3$ for most of case scenarios. From Figure 8(a)-(b), we can see that the choice of the bandwidth parameter ϕ has a great impacts on the dGCV* score as well as the penalty parameter λ. It seems that the latter provides some additional small adjustments after a good value of ϕ is chosen.

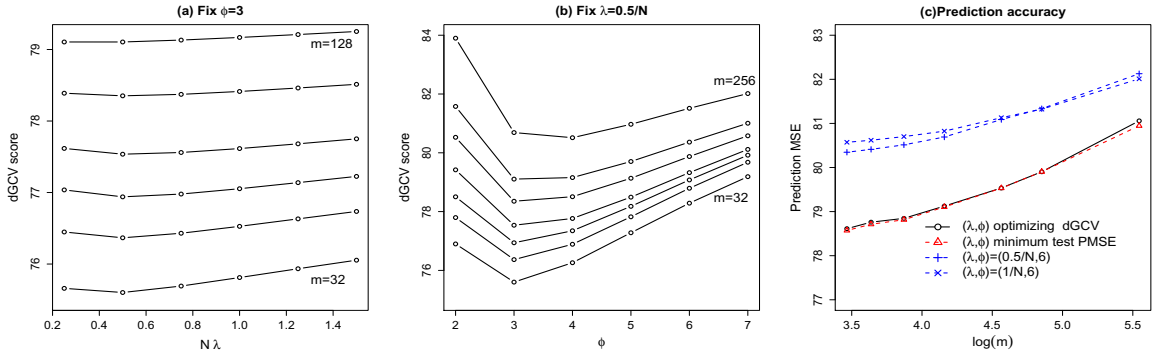

Figure 8: (a) dGCV score v.s. $N\lambda$ with $m = 32$ (the bottommost) to $m = 128$ (the uppermost); (b) dGCV score v.s. ϕ with $m = 32$ (the bottommost) to $m = 256$ (the uppermost); (c) The prediction mean squared errors on the testing samples v.s. $\log(m)$.

In Zhang et al. (2015), the authors used a fixed value $\lambda = 1/N$ and a $\phi = 6$ chosen by the cross-validation for their kernel ridge regression model. In Figure 8(c), we can see that such a choice leads to a much worse prediction mean squared error (PMSE) on the testing samples. Using the proposed dGCV criterion, our choice of λ and ϕ yields almost identical prediction accuracy as the minimum possible PMSE on the testing samples obtained over all 36 grid points.

Remark 3. *Note that for any given combination of* (λ, ϕ)*, the estimated function* $\bar{f}_{\lambda,\phi}$ *used in* dGCV* *is the same for different values of* m^**, which is defined in* (3)*. The agreement between the test PMSE of the* dGCV* *method and the minimum test PMSE in Figure 8(c) suggests that there is no room to improve over the predictive performance of* $\bar{f}_{\lambda,\phi}$ *using tuning parameters selected by* dGCV**, as long as the same multivariate Gaussian reproducing kernel function is used. This is a strong indication that* $m^* = \lceil m/10 \rceil$ *is a good choice for this example, considering that* dGCV* *did not use any information of the* $51,630$ *testing examples.*

7 Discussion

In this paper, we proposed a data-driven criterion named dGCV that can be used to empirically selecting the critical tuning parameter λ for d-KRR. Not only the proposed approach is computationally scalable even for massive data sets, we have also theoretically shown that it is asymptotically optimal in the sense that minimizing dGCV

is equivalent to minimizing the true global conditional empirical loss, extending the existing optimality results of GCV to the divide-and-conquer framework.

There are a few ways to extend the current work. For example, we have so far presumed a fixed m. One important direction is to investigate the growth rate of m for some specific kernels under which Theorem 1 still holds, following the framework proposed in Shang and Cheng (2017). It is also of practical interest to develop a justifiable data-driven approach to detect the breaking point for m. Another interesting research direction is to develop a tuning criterion similar to the dGCV for more general panelized Kernel regression such as Zhang et al. (2016) and Chen et al. (2017). The definition of dGCV in (16) relies heavily on the closed form solution to the Kernel ridge regression, which is not available if the loss function or the penalty in (2) are replaced by the quantile loss or the lasso penalty, respectively. The major difficulty lies in how to replace the effective degrees of freedom $\mathrm{tr}\{\mathbf{A}_{kk}(\lambda)\}$'s in the denominator of (16) when the hat matrices $\mathbf{A}_{kk}$'s do not exist. Although there has been some research on this issue such as Yuan (2006), much more thorough investigations are needed.

Acknowledgments

Ganggang Xu's research is partially supported by Collaboration Grants for Mathematicians (Award ID: 524205) from Simons Foundation and NSF Award SES-1902195. Zuofeng Shang's research is supported by NSF Award DMS-1764280 and DMS-1821157. Guang Cheng's research is partially supported by NSF CAREER Award DMS-1151692, DMS1418042, DMS-1712907, DMS-1811812 and Office of Naval Research (ONR N00014-15-1-2331, ONR N00014-18-1-2759).

References

Abramowitz, M., Stegun, I. A., et al. (1972), *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, vol. 55, Dover publications New York.

Bach, F. (2013), "Sharp analysis of low-rank kernel matrix approximations," in *Conference on Learning Theory*, pp. 185–209.

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011), "The Million Song Dataset." in *Ismir*, vol. 2, p. 10.

Blanchard, G. and Mücke, N. (2016), "Parallelizing spectral algorithms for kernel learning," *arXiv preprint arXiv:1610.07487*.

Chang, X., Lin, S.-B., and Zhou, D.-X. (2017), "Distributed semi-supervised learning with kernel ridge regression," *Journal of Machine Learning Research*, 18, 1–22.

Chen, J., Zhang, C., Kosorok, M. R., and Liu, Y. (2017), "Double Sparsity Kernel Learning with Automatic Variable Selection and Data Extraction," *arXiv preprint arXiv:1706.01426*.

Chen, X. and Xie, M. (2014), "A split-and-conquer approach for analysis of extraordinarily large data," *Statistica Sinica*, 24, 1655–1684.

Craven, P. and Wahba, G. (1978), "Smoothing noisy data with spline functions," *Numerische mathematik*, 31, 377–403.

Gu, C. (2013), *Smoothing spline ANOVA models*, vol. 297, Springer Science & Business Media.

Gu, C. and Ma, P. (2005), "Optimal smoothing in nonparametric mixed-effect models," *The Annals of Statistics*, 33, 1357–1379.

Guo, Z.-C., Shi, L., and Wu, Q. (2017), "Learning theory of distributed regression with bias corrected regularization kernel network," *Journal of Machine Learning Research*, 18, 1–25.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006), *A distribution-free theory of nonparametric regression*, Springer Science & Business Media.

Hastie, T. J. (2017), "Generalized additive models," in *Statistical models in S*, Routledge, pp. 249–307.

Jehan, T. and DesRoches, D. (2011), "Analyzer documentation," *The Echo Nest*.

Kandasamy, K. and Yu, Y. (2016), "Additive approximations in high dimensional nonparametric regression via the SALSA," in *International Conference on Machine Learning*, pp. 69–78.

Li, K.-C. (1986), "Asymptotic optimality of CL and generalized cross-validation in ridge regression with application to spline smoothing," *The Annals of Statistics*, 14, 1101–1112.

Lin, S.-B., Guo, X., and Zhou, D.-X. (2017), "Distributed learning with regularized least squares," *The Journal of Machine Learning Research*, 18, 3202–3232.

Lu, J., Cheng, G., and Liu, H. (2016), "Nonparametric Heterogeneity Testing For Massive Data," *arXiv preprint arXiv:1601.06212*.

Mallows, C. L. (2000), "Some comments on Cp," *Technometrics*, 42, 87–94.

Pollard, D. (1986), "Rates of uniform almost-sure convergence for empirical processes indexed by unbounded classes of functions," .

— (1995), "Uniform ratio limit theorems for empirical processes," *Scandinavian Journal of Statistics*, 271–278.

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rice, J. and Rosenblatt, M. (1983), "Smoothing splines: regression, derivatives and deconvolution," *The annals of Statistics*, 141–156.

Schaback, R. and Wendland, H. (2006), "Kernel techniques: from machine learning to meshless methods," *Acta numerica*, 15, 543–639.

Shang, Z. and Cheng, G. (2017), "Computational limits of a distributed algorithm for smoothing spline," *The Journal of Machine Learning Research*, 18, 3809–3845.

Shawe-Taylor, J. and Cristianini, N. (2004), *Kernel methods for pattern analysis*, Cambridge university press.

van de Geer, S. A. and van de Geer, S. (2000), *Empirical Processes in M-estimation*, vol. 6, Cambridge university press.

Wahba, G. (1990), *Spline models for observational data*, vol. 59, Siam.

Wood, S. N. (2004), "Stable and efficient multiple smoothing parameter estimation for generalized additive models," *Journal of the American Statistical Association*, 99, 673–686.

Xiang, D. and Wahba, G. (1996), "A generalized approximate cross validation for smoothing splines with non-Gaussian data," *Statistica Sinica*, 6, 675–692.

Xu, G. and Huang, J. Z. (2012), "Asymptotic optimality and efficient computation of the leave-subject-out cross-validation," *The Annals of Statistics*, 40, 3003–3030.

Xu, G., Shang, Z., and Cheng, G. (2018), "Optimal tuning for divide-and-conquer kernel ridge regression with massive data," in *Proceedings of the 35th International Conference on Machine Learning, PMLR*, vol. 80, pp. 5483–5491.

Yuan, M. (2006), "GACV for quantile smoothing splines," *Computational statistics & data analysis*, 50, 813–829.

Zhang, C., Liu, Y., and Wu, Y. (2016), "On quantile regression in reproducing kernel Hilbert spaces with the data sparsity constraint," *The Journal of Machine Learning Research*, 17, 1374–1418.

Zhang, Y., Duchi, J., and Wainwright, M. (2015), "Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates," *The Journal of Machine Learning Research*, 16, 3299–3340.

Zhao, T., Cheng, G., and Liu, H. (2016), "A partially linear framework for massive heterogeneous data," *Annals of statistics*, 44, 1400.

Appendix

From now on, we suppress the dependence of $\mathbf{A}_{kl}(\lambda)$'s and $\bar{\mathbf{A}}(\lambda)$ on λ for ease of presentation and simply use $\mathbf{A}_{kl}$'s and $\bar{\mathbf{A}}$ whenever there is no ambiguity.

Lemma A.1. *Under the condition C1, we have that* $\lambda_{max}(\bar{\mathbf{A}}_m\bar{\mathbf{A}}_m^T) = O_{\mathbb{P}_X}(1)$.

Proof. Define the following matrix

$$\bar{\mathbf{K}}_m = \frac{1}{m}\begin{pmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} & \cdots & \mathbf{K}_{1m} \\ \mathbf{K}_{21} & \mathbf{K}_{22} & \cdots & \mathbf{K}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{K}_{m1} & \mathbf{K}_{m2} & \cdots & \mathbf{K}_{mm} \end{pmatrix}.$$

Then it is straightforward to see that

$$\bar{\mathbf{A}}_m\bar{\mathbf{A}}_m^T = \bar{\mathbf{K}}\mathbf{D}_1\bar{\mathbf{K}}^T,$$

where $\mathbf{D}_1 = \text{diag}\{\mathbf{B}_{11}, \ldots, \mathbf{B}_{mm}\}$ with $\mathbf{B}_{ll} = (\mathbf{K}_{ll} + n_l\lambda\mathbf{I}_l)^{-2}$, for $l = 1, \ldots, m$. Then

$$\bar{\mathbf{K}}\mathbf{D}_1\bar{\mathbf{K}}^T = \frac{1}{m^2}\begin{pmatrix} \mathbf{K}_{11} \\ \mathbf{K}_{21} \\ \vdots \\ \mathbf{K}_{m1} \end{pmatrix}\mathbf{B}_{11}(\mathbf{K}_{11}^T, \ldots, \mathbf{K}_{m1}^T) + \cdots + \frac{1}{m^2}\begin{pmatrix} \mathbf{K}_{1m} \\ \mathbf{K}_{2m} \\ \vdots \\ \mathbf{K}_{mm} \end{pmatrix}\mathbf{B}_{mm}(\mathbf{K}_{1m}^T, \ldots, \mathbf{K}_{mm}^T),$$

which implies that

$$\lambda_{\max}(\bar{\mathbf{A}}_m\bar{\mathbf{A}}_m^T) \leq \frac{1}{m^2}\sum_{l=1}^m \lambda_{\max}\Big\{\begin{pmatrix} \mathbf{K}_{1l} \\ \mathbf{K}_{2l} \\ \vdots \\ \mathbf{K}_{ml} \end{pmatrix}\mathbf{B}_{ll}(\mathbf{K}_{1l}^T, \ldots, \mathbf{K}_{ml}^T)\Big\} = \frac{1}{m^2}\sum_{l=1}^m \lambda_{\max}(\mathbf{B}_{ll}\sum_{k=1}^m \mathbf{K}_{kl}^T\mathbf{K}_{kl})$$

$$= \frac{1}{m}\sum_{l=1}^m \lambda_{\max}\left\{(\mathbf{K}_{ll} + n_l\lambda\mathbf{I}_l)^{-2}\left(\frac{1}{m}\sum_{k=1}^m \mathbf{K}_{kl}^T\mathbf{K}_{kl}\right)\right\} = O_{\mathbb{P}_X}(1).$$

The last inequality follows from condition C1. □

Lemma A.2. *Under the conditions C1-C2 and C3(a), for a fixed* λ, *we have that*

$$\bar{L}(\lambda|\mathbf{X}) - \bar{R}(\lambda|\mathbf{X}) = o_{\mathbb{P}_{\varepsilon,X}}\{\bar{R}(\lambda|\mathbf{X})\}. \tag{A.1}$$

Proof. Using similar notations in equation (14), it is straightforward to show that

$$\bar{L}(\lambda|\mathbf{X}) = \frac{1}{N}\left(\bar{\mathbf{A}}_m\mathbf{Y} - \boldsymbol{F}\right)^T\mathbf{W}\left(\bar{\mathbf{A}}_m\mathbf{Y} - \boldsymbol{F}\right), \text{ with } \mathbf{Y} = \boldsymbol{F} + \boldsymbol{\varepsilon}. \tag{A.2}$$

Using (14), we have that

$$\bar{L}(\lambda|\mathbf{X}) - \bar{R}(\lambda|\mathbf{X}) = -\frac{2}{N}\boldsymbol{F}^T(\mathbf{I} - \bar{\mathbf{A}}_m)^T\mathbf{W}\bar{\mathbf{A}}_m\boldsymbol{\varepsilon} + \frac{1}{N}\boldsymbol{\varepsilon}^T\bar{\mathbf{A}}_m^T\mathbf{W}\bar{\mathbf{A}}_m\boldsymbol{\varepsilon} - \frac{\sigma^2}{N}\mathrm{tr}(\bar{\mathbf{A}}_m^T\mathbf{W}\bar{\mathbf{A}}_m).$$

Since the random error ε and the covariate X are independent in model (1), to show (A.1), it suffices to show the following two equations

$$\mathrm{Var}_\varepsilon\left\{\frac{1}{N}\boldsymbol{F}^T(\mathbf{I} - \bar{\mathbf{A}}_m)^T\mathbf{W}\bar{\mathbf{A}}_m\boldsymbol{\varepsilon}\right\} = o_{\mathbb{P}_X}\{\bar{R}^2(\lambda|\mathbf{X})\}, \tag{A.3}$$

$$\mathrm{Var}_\varepsilon\left\{\frac{1}{N}\boldsymbol{\varepsilon}^T\bar{\mathbf{A}}_m^T\mathbf{W}\bar{\mathbf{A}}_m\boldsymbol{\varepsilon} - \frac{\sigma^2}{N}\mathrm{tr}(\bar{\mathbf{A}}_m^T\mathbf{W}\bar{\mathbf{A}}_m)\right\} = o_{\mathbb{P}_X}\{\bar{R}^2(\lambda|\mathbf{X})\}. \tag{A.4}$$

We first show (A.3). Straightforward algebra yields that

$$\begin{aligned}
\mathrm{Var}_\varepsilon\left\{\frac{1}{N}\boldsymbol{F}^T(\mathbf{I} - \bar{\mathbf{A}}_m)^T\mathbf{W}\bar{\mathbf{A}}_m\boldsymbol{\varepsilon}\right\} &= \frac{\sigma^2}{N^2}\boldsymbol{F}^T(\mathbf{I} - \bar{\mathbf{A}}_m)^T\mathbf{W}\left(\bar{\mathbf{A}}_m\bar{\mathbf{A}}_m^T\right)\mathbf{W}(\mathbf{I} - \bar{\mathbf{A}}_m)\boldsymbol{F} \\
&\le \frac{\sigma^2\lambda_{\max}\left(\bar{\mathbf{A}}_m\bar{\mathbf{A}}_m^T\mathbf{W}\right)}{N}\frac{1}{N}\boldsymbol{F}^T(\mathbf{I} - \bar{\mathbf{A}}_m)^T\mathbf{W}(\mathbf{I} - \bar{\mathbf{A}}_m)\boldsymbol{F} \\
&\le \frac{\sigma^2\lambda_{\max}\left(\bar{\mathbf{A}}_m\bar{\mathbf{A}}_m^T\right)\lambda_{\max}(\mathbf{W})}{N\bar{R}(\lambda|\mathbf{X})}\bar{R}^2(\lambda|\mathbf{X}) \\
&= o_{\mathbb{P}_X}(1)\bar{R}^2(\lambda|\mathbf{X}) = o_{\mathbb{P}_X}\{\bar{R}^2(\lambda|\mathbf{X})\},
\end{aligned}$$

where the second last equation follows from conditions C2 and C3(a) and Lemma (A.1).

Now we show (A.4). Straightforward algebra yields that

$$\begin{aligned}
&\mathrm{Var}_\varepsilon\left\{\frac{1}{N}\boldsymbol{\varepsilon}^T\bar{\mathbf{A}}_m^T\mathbf{W}\bar{\mathbf{A}}_m\boldsymbol{\varepsilon} - \frac{\sigma^2}{N}\mathrm{tr}(\bar{\mathbf{A}}_m^T\mathbf{W}\bar{\mathbf{A}}_m)\right\} = \frac{\mathbb{E}_\varepsilon\varepsilon^4 - \sigma^4}{N^2}\sum_{i=1}^N\bar{b}_{ii}^2 + 2\sigma^4\sum_i\sum_{j}^{i\neq j}b_{ij}^2 \\
&\le \frac{K_1}{N^2}\mathrm{tr}\{(\bar{\mathbf{A}}_m^T\mathbf{W}\bar{\mathbf{A}}_m)^2\} \le \frac{K_1\lambda_{\max}(\bar{\mathbf{A}}_m^T\mathbf{W}\bar{\mathbf{A}}_m)}{N^2}\mathrm{tr}(\bar{\mathbf{A}}_m^T\mathbf{W}\bar{\mathbf{A}}_m) \\
&\le \frac{K_1\lambda_{\max}(\bar{\mathbf{A}}_m^T\mathbf{W}\bar{\mathbf{A}}_m)}{N\sigma^2}\bar{R}(\lambda|\mathbf{X}) \le \frac{K_1\lambda_{\max}(\bar{\mathbf{A}}_m^T\bar{\mathbf{A}}_m)\lambda_{\max}(\mathbf{W})}{\sigma^2N\bar{R}(\lambda|\mathbf{X})}\bar{R}^2(\lambda|\mathbf{X}) \\
&= o_{\mathbb{P}_X}(1)\bar{R}^2(\lambda|\mathbf{X})
\end{aligned} \tag{A.5}$$

where $\bar{b}_{ij}$ is the (i,j)th element of matrix $\bar{\mathbf{A}}_m^T\mathbf{W}\bar{\mathbf{A}}_m$ and $K_1 = \mathbb{E}_\varepsilon\varepsilon^4 + \sigma^4$. The last equality follows from conditions C2 and C3(a) and Lemma A.1. Using (A.3)-(A.4), the equation (A.1) follows from a simple application of the Cauchy-Schwartz inequality and the Markov's inequality. The proof is complete. □

Proof of Lemma 1. Using (A.2) and (15), we have that

$$\bar{U}(\lambda|\mathbf{X}) - \bar{L}(\lambda|\mathbf{X}) - \frac{1}{N}\boldsymbol{\varepsilon}^T\mathbf{W}\boldsymbol{\varepsilon} = \frac{2}{N}\boldsymbol{F}^T(\mathbf{I} - \bar{\mathbf{A}}_m)^T\mathbf{W}\boldsymbol{\varepsilon} - \frac{2}{N}\left\{\boldsymbol{\varepsilon}^T\bar{\mathbf{A}}_m\mathbf{W}\boldsymbol{\varepsilon} - \sigma^2\mathrm{tr}(\bar{\mathbf{A}}_m\mathbf{W})\right\}. \tag{A.6}$$

Notice that the random error ε and the covariate X are independent in model (1). We will show (18) using equation (A.1) in Lemma A.2, for which it suffices to show the following two equations

$$\mathrm{Var}_\varepsilon\left\{\frac{1}{N}\boldsymbol{F}^T(\mathbf{I}-\bar{\mathbf{A}}_m)^T\mathbf{W}\boldsymbol{\varepsilon}\right\} = o_{\mathbb{P}_X}\{\bar{R}^2(\lambda|\mathbf{X})\}, \tag{A.7}$$

$$\mathrm{Var}_\varepsilon\left\{\frac{1}{N}\boldsymbol{\varepsilon}^T\bar{\mathbf{A}}_m\mathbf{W}\boldsymbol{\varepsilon}-\frac{\sigma^2}{N}\mathrm{tr}(\bar{\mathbf{A}}_m\mathbf{W})\right\} = o_{\mathbb{P}_X}\{\bar{R}^2(\lambda|\mathbf{X})\}. \tag{A.8}$$

We first show (A.7). Straightforward algebra yields that

$$\begin{aligned}\mathrm{Var}_\varepsilon\left\{\frac{1}{N}\boldsymbol{F}^T(\mathbf{I}-\bar{\mathbf{A}}_m)^T\mathbf{W}\boldsymbol{\varepsilon}\right\} &= \frac{\sigma^2}{N^2}\boldsymbol{F}^T(\mathbf{I}-\bar{\mathbf{A}}_m)^T\mathbf{W}^2(\mathbf{I}-\bar{\mathbf{A}}_m)\boldsymbol{F} \le \frac{\sigma^2\lambda_{\max}(\mathbf{W})}{N\bar{R}(\lambda|\mathbf{X})}\bar{R}^2(\lambda|\mathbf{X})\\ &= o_{\mathbb{P}_X}(1)\bar{R}^2(\lambda|\mathbf{X}) = o_{\mathbb{P}_X}\{\bar{R}^2(\lambda|\mathbf{X})\},\end{aligned}$$

where the second last equation follows from conditions C2-C3. Next, we show (A.8). Using condition C2, similar to the inequality (A.5), it is straightforward to show that

$$\begin{aligned}\mathrm{Var}_\varepsilon\left\{\frac{1}{N}\boldsymbol{\varepsilon}^T\bar{\mathbf{A}}_m\mathbf{W}\boldsymbol{\varepsilon}\right\} &\le \frac{K_1}{N^2}\mathrm{tr}(\bar{\mathbf{A}}_m^T\mathbf{W}^2\bar{\mathbf{A}}_m) \le \frac{K_1\lambda_{\max}(\mathbf{W})}{N\sigma^2}\bar{R}(\lambda|\mathbf{X})\\ &= \frac{K_1\lambda_{\max}(\mathbf{W})}{\sigma^2N\bar{R}(\lambda|\mathbf{X})}\bar{R}^2(\lambda|\mathbf{X}) = o_{\mathbb{P}_X}(1)\bar{R}^2(\lambda|\mathbf{X}),\end{aligned}$$

where $K_1 = \mathbb{E}_\varepsilon\varepsilon^4+\sigma^4$ is bounded. Hence, (A.8) is proved using, again, condition C2-C3. Using (A.7)-(A.8) and (A.1), the equation (18) follows from a simple application of the Cauchy-Schwartz inequality and the Markov's inequality. The proof is complete. □

Proof of Theorem 1 . Using Lemma 1 and Lemma A.2, it suffices to show that

$$\mathrm{dGCV}_{DC}(\lambda|\mathbf{X}) - \bar{U}(\lambda|\mathbf{X}) = o_{\mathbb{P}_{\varepsilon,X}}\{\bar{R}(\lambda|\mathbf{X})\}. \tag{A.9}$$

Using the first order Taylor expansion of $(1-x)^{-2}$ around $x=0$, we have that $(1-x)^{-2} = 1+2x+3(1-x^*)^{-4}x^2$ for some $x^*\in(0,x)$. Under condition C3, we have that $\frac{\mathrm{tr}(\bar{\mathbf{A}}_m)}{N} = o_{\mathbb{P}_X}(1)$ and thus we can consider the following decomposition

$$\begin{aligned}\mathrm{dGCV}(\lambda|\mathbf{X}) - \bar{U}(\lambda|\mathbf{X}) = &\underbrace{\left\{\frac{1}{N}\mathbf{Y}^T\{\mathbf{I}-\bar{\mathbf{A}}_m(\lambda)\}^T\mathbf{W}\{\mathbf{I}-\bar{\mathbf{A}}_m(\lambda)\}\mathbf{Y}-\sigma^2\right\}\frac{2\mathrm{tr}(\bar{\mathbf{A}}_m\mathbf{W})}{N}}_{I}\\ &+\underbrace{\frac{1}{N}\mathbf{Y}^T\{\mathbf{I}-\bar{\mathbf{A}}_m(\lambda)\}^T\mathbf{W}\{\mathbf{I}-\bar{\mathbf{A}}_m(\lambda)\}\mathbf{Y}O_{\mathbb{P}_X}\left(\frac{\{\mathrm{tr}(\bar{\mathbf{A}}_m\mathbf{W})\}^2}{N^2}\right)}_{II}\end{aligned}$$

Using condition C4, we have that

$$\frac{\mathrm{tr}(\bar{\mathbf{A}}_m\mathbf{W})}{N} = o_{\mathbb{P}_X}\{\bar{R}^{1/2}(\lambda|\mathbf{X})\}, \tag{A.10}$$

which implies that $II = o_{\mathbb{P}_X}(\bar{R}(\lambda|\mathbf{X}))$ since $\frac{1}{N}\mathbf{Y}^T\{\mathbf{I}-\bar{\mathbf{A}}_m(\lambda)\}^T\mathbf{W}\{\mathbf{I}-\bar{\mathbf{A}}_m(\lambda)\}\mathbf{Y}$ is

bounded. For part I, we can write

$$I = \left\{ \frac{1}{N}\mathbf{Y}^T\{\mathbf{I} - \bar{\mathbf{A}}_m(\lambda)\}^T\mathbf{W}\{\mathbf{I} - \bar{\mathbf{A}}_m(\lambda)\}\mathbf{Y} - \sigma^2 \right\} \frac{2\text{tr}(\bar{\mathbf{A}}_m\mathbf{W})}{N}$$
$$= \left\{ \bar{U}(\lambda|\mathbf{X}) - \frac{1}{N}\boldsymbol{\varepsilon}^T\mathbf{W}\boldsymbol{\varepsilon} \right\} \frac{2\text{tr}(\bar{\mathbf{A}}_m\mathbf{W})}{N} + \left(\frac{1}{N}\boldsymbol{\varepsilon}^T\mathbf{W}\boldsymbol{\varepsilon} - \sigma^2 \right) \frac{2\text{tr}(\bar{\mathbf{A}}_m\mathbf{W})}{N} - \frac{4\{\text{tr}(\bar{\mathbf{A}}_m\mathbf{W})\}^2\sigma^2}{N^2}.$$

By Lemma 1, we have that $\bar{U}(\lambda|\mathbf{X}) - \frac{1}{N}\boldsymbol{\varepsilon}^T\mathbf{W}\boldsymbol{\varepsilon} = \bar{R}(\lambda|\mathbf{X}) + o_{\mathbb{P}_{\varepsilon,X}}\{\bar{R}(\lambda|\mathbf{X})\}$. Under condition C3, one has that $\frac{\text{tr}(\bar{\mathbf{A}}_m\mathbf{W})}{N} = o_{\mathbb{P}_X}(1)$, and thus

$$\left\{ \bar{U}(\lambda|\mathbf{X}) - \frac{1}{N}\boldsymbol{\varepsilon}^T\mathbf{W}\boldsymbol{\varepsilon} \right\} \frac{2\text{tr}(\bar{\mathbf{A}}_m\mathbf{W})}{N} = o_{\mathbb{P}_{\varepsilon,X}}\{\bar{R}(\lambda|\mathbf{X})\}.$$

Furthermore, since $\frac{1}{N}\boldsymbol{\varepsilon}^T\mathbf{W}\boldsymbol{\varepsilon} - \sigma^2 = O_{\mathbb{P}_\varepsilon}(N^{-1/2})$ (condition C3 (a)) and $N\bar{R}(\lambda|\mathbf{X}) \xrightarrow{\mathbb{P}_X} \infty$ (condition C2), we have that $\frac{1}{N}\boldsymbol{\varepsilon}^T\mathbf{W}\boldsymbol{\varepsilon} - \sigma^2 = o_{\mathbb{P}_{\varepsilon,X}}\{\bar{R}^{1/2}(\lambda|\mathbf{X})\}$. Using this and equation (A.10), we have that

$$\left(\frac{1}{N}\boldsymbol{\varepsilon}^T\mathbf{W}\boldsymbol{\varepsilon} - \sigma^2 \right) \frac{2\text{tr}(\bar{\mathbf{A}}_m\mathbf{W})}{N} = o_{\mathbb{P}_{\varepsilon,X}}\{\bar{R}(\lambda|\mathbf{X})\}.$$

The third part of I is $o_{\mathbb{P}_X}\{\bar{R}(\lambda|\mathbf{X})\}$ due to equation (A.10). Therefore, we have shown that

$$\text{dGCV}(\lambda|\mathbf{X}) - \bar{U}(\lambda|\mathbf{X}) = o_{\mathbb{P}_{\varepsilon,X}}\{\bar{R}(\lambda|\mathbf{X})\},$$

which completes the proof. □

Lemma A.3. *Define the following class of non-negative functions*

$$\mathcal{F} = \{f \in L_2(\mathbb{P}) : f \geq 0, \|f\|_{\text{sup}} \leq V, J_1(f) \leq V^2H^2\}, \tag{A.11}$$

where $V > 0$ and $H > 0$ are constants. If condition C4'(d) holds and (ϵ_n, γ_n) satisfy

$$\epsilon_n^3\gamma_n^2 \geq \frac{c_0(1+H)V}{n}, \tag{A.12}$$

where $c_0 > 0$ is a constant, then there exists a constant $C > 0$ such that for all n,

$$P\left(\sup_{f\in\mathcal{F}} \frac{|\mathbb{P}_n f - \mathbb{P}f|}{\mathbb{P}_n f + \mathbb{P}f + \gamma_n(\mathbb{P}_n f + \mathbb{P}f + 1)} > C\epsilon_n \right) \leq \exp(-n\epsilon_n^2\gamma_n/2).$$

Proof. Recall the definition of $\mathcal{F}_0$ in condition C4'(d). It can be checked that

$$\mathcal{F} \subseteq V(1+H)\mathcal{F}_0.$$

Hence under condition C4'(d), we have that with probability approaching one,

$$\begin{aligned} N(\epsilon_n\gamma_n, \|\cdot\|_{\mathbb{P}_n}, \mathcal{F}) &\leq N(\epsilon_n\gamma_n, \|\cdot\|_{\mathbb{P}_n}, V(1+H)\mathcal{F}_0) = N\left(\frac{\epsilon_n\gamma_n}{V(1+H)}, \|\cdot\|_{\mathbb{P}_n}, \mathcal{F}_0 \right) \\ &\leq \exp\left\{ \frac{C_0(1+H)V}{\epsilon_n\gamma_n} \right\}. \end{aligned}$$

By the Theorem given in Pollard (1995) and the Theorem 2.1 of Pollard (1986), there exists constants C and c_0 such that

$$\begin{aligned} P\left(\sup_{f\in\mathcal{F}}\frac{|\mathbb{P}_n f-\mathbb{P}f|}{\mathbb{P}_n f+\mathbb{P}f+\gamma_n(\mathbb{P}_n f+\mathbb{P}f+1)}>C\epsilon_n\right) &\leq \exp\left(c_0\frac{(1+H)V}{2\epsilon_n\gamma_n}-n\epsilon_n^2\gamma_n\right) \\ &\leq \exp(-n\epsilon_n^2\gamma_n/2). \end{aligned}$$

□

Proof of Lemma 2. Define the kernel matrix

$$\mathbf{K}=\begin{pmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} & \cdots & \mathbf{K}_{1m} \\ \mathbf{K}_{21} & \mathbf{K}_{22} & \cdots & \mathbf{K}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{K}_{m1} & \mathbf{K}_{m2} & \cdots & \mathbf{K}_{mm} \end{pmatrix}=\mathbf{\Phi}\mathbf{\Phi}^T,$$

where $\mathbf{\Phi}$ is a $N\times r$ matrix with r being the rank of $\mathbf{K}$. By this notation, we have that

$$\mathbf{K}_{ll}=\mathbf{\Phi}_l\mathbf{\Phi}_l^T, \qquad l=1,\cdots,m.$$

where $\mathbf{\Phi}_l$ is a $n_l\times r$ submatrix of $\mathbf{\Phi}$ consists of rows corresponding to a subdata set S_l. Then it is straightforward to show that

$$\lambda_{\max}\left\{(\mathbf{K}_{ll}+n_l\lambda\mathbf{I}_l)^{-2}\left(\frac{1}{m}\sum_{k=1}^m\mathbf{K}_{kl}^T\mathbf{K}_{kl}\right)\right\}=\frac{1}{m}\lambda_{\max}\left\{\mathbf{\Phi}\mathbf{\Phi}_l^T(\mathbf{\Phi}_l\mathbf{\Phi}_l^T+n_l\lambda\mathbf{I}_l)^{-2}\mathbf{\Phi}_l\mathbf{\Phi}^T\right\}.$$

Using the Sherman–Morrison formula, we can show that

$$\begin{aligned} \mathbf{\Phi}_l^T(\mathbf{\Phi}_l\mathbf{\Phi}_l^T+n_l\lambda\mathbf{I}_l)^{-1} &= \mathbf{\Phi}_l^T\left[n^{-1}\lambda^{-1}\mathbf{I}-n^{-2}\lambda^{-2}\mathbf{\Phi}_l(\mathbf{I}+n^{-1}\lambda^{-1}\mathbf{\Phi}_l^T\mathbf{\Phi}_l)^{-1}\mathbf{\Phi}_l^T\right] \\ &= n^{-1}\lambda^{-1}\left[\mathbf{I}-n^{-1}\lambda^{-1}\mathbf{\Phi}_l^T\mathbf{\Phi}_l(\mathbf{I}+n^{-1}\lambda^{-1}\mathbf{\Phi}_l^T\mathbf{\Phi}_l)^{-1}\right]\mathbf{\Phi}_l^T \\ &= n^{-1}\lambda^{-1}\left[(\mathbf{I}+n^{-1}\lambda^{-1}\mathbf{\Phi}_l^T\mathbf{\Phi}_l)^{-1}\right]\mathbf{\Phi}_l^T, \end{aligned}$$

which gives that

$$\begin{aligned} &\lambda_{\max}\left\{(\mathbf{K}_{ll}+n_l\lambda\mathbf{I}_l)^{-2}\left(\frac{1}{m}\sum_{k=1}^m\mathbf{K}_{kl}^T\mathbf{K}_{kl}\right)\right\} \\ &=\frac{n^{-2}\lambda^{-2}}{m}\lambda_{\max}\left\{\mathbf{\Phi}\left[(\mathbf{I}+n^{-1}\lambda^{-1}\mathbf{\Phi}_l^T\mathbf{\Phi}_l)^{-1}\right]\mathbf{\Phi}_l^T\mathbf{\Phi}_l\left[(\mathbf{I}+n^{-1}\lambda^{-1}\mathbf{\Phi}_l^T\mathbf{\Phi}_l)^{-1}\right]\mathbf{\Phi}^T\right\} \\ &=\frac{n^{-1}\lambda^{-1}}{m}\lambda_{\max}\left\{\mathbf{\Phi}(\mathbf{I}+n^{-1}\lambda^{-1}\mathbf{\Phi}_l^T\mathbf{\Phi}_l)^{-1}\mathbf{\Phi}^T-\mathbf{\Phi}(\mathbf{I}+n^{-1}\lambda^{-1}\mathbf{\Phi}_l^T\mathbf{\Phi}_l)^{-2}\mathbf{\Phi}^T\right\} \\ &\leq\frac{1}{N}\lambda_{\max}\left\{\mathbf{\Phi}(\lambda\mathbf{I}+n^{-1}\mathbf{\Phi}_l^T\mathbf{\Phi}_l)^{-1}\mathbf{\Phi}^T\right\}=\lambda_{\max}\left\{(\lambda\mathbf{I}+n^{-1}\mathbf{\Phi}_l^T\mathbf{\Phi}_l)^{-1}\left[\frac{1}{N}\mathbf{\Phi}^T\mathbf{\Phi}\right]\right\}. \end{aligned}$$

Using the following identity from the Appendix B of Bach (2013)

$$(\lambda \mathbf{I} + n^{-1}\mathbf{\Phi}_l^T\mathbf{\Phi}_l)^{-1} = \left(\lambda \mathbf{I} + \frac{1}{N}\mathbf{\Phi}^T\mathbf{\Phi} - \frac{1}{N}\mathbf{\Phi}^T\mathbf{\Phi} + n^{-1}\mathbf{\Phi}_l^T\mathbf{\Phi}_l\right)^{-1}$$
$$= \left(\lambda \mathbf{I} + \frac{1}{N}\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1/2}\left[\mathbf{I} - \frac{1}{N}\mathbf{\Psi}^T\mathbf{\Psi} + \frac{1}{n}\mathbf{\Psi}_l^T\mathbf{\Psi}_l\right]^{-1}\left(\lambda \mathbf{I} + \frac{1}{N}\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1/2},$$

where $\mathbf{\Psi} = \mathbf{\Phi}\left(\lambda \mathbf{I} + \frac{1}{N}\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1/2}$ and $\mathbf{\Psi}_l$ is the submatrix of $\mathbf{\Psi}$, we have that

$$\lambda_{\max}\left\{(\mathbf{K}_{ll} + n_l\lambda\mathbf{I}_l)^{-2}\left(\frac{1}{m}\sum_{k=1}^{m}\mathbf{K}_{kl}^T\mathbf{K}_{kl}\right)\right\} \leq \lambda_{\max}\left\{(\lambda \mathbf{I} + n^{-1}\mathbf{\Phi}_l^T\mathbf{\Phi}_l)^{-1}\left[\frac{1}{N}\mathbf{\Phi}^T\mathbf{\Phi}\right]\right\}$$
$$= \lambda_{\max}\left\{\left(\lambda \mathbf{I} + \frac{1}{N}\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1/2}\left[\mathbf{I} - \frac{1}{N}\mathbf{\Psi}^T\mathbf{\Psi} + \frac{1}{n}\mathbf{\Psi}_l^T\mathbf{\Psi}_l\right]^{-1}\left(\lambda \mathbf{I} + \frac{1}{N}\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1/2}\left[\frac{1}{N}\mathbf{\Phi}^T\mathbf{\Phi}\right]\right\}$$
$$\leq \sigma_{\max}\left\{\left[\mathbf{I} - \frac{1}{N}\mathbf{\Psi}^T\mathbf{\Psi} + \frac{1}{n}\mathbf{\Psi}_l^T\mathbf{\Psi}_l\right]^{-1}\right\}\lambda_{\max}\left\{\left(\lambda \mathbf{I} + \frac{1}{N}\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1/2}\left[\frac{1}{N}\mathbf{\Phi}^T\mathbf{\Phi}\right]\left(\lambda \mathbf{I} + \frac{1}{N}\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1/2}\right\}$$
$$\leq \sigma_{\max}\left\{\left[\mathbf{I} - \frac{1}{N}\mathbf{\Psi}^T\mathbf{\Psi} + \frac{1}{n}\mathbf{\Psi}_l^T\mathbf{\Psi}_l\right]^{-1}\right\},$$

where $\sigma_{\max}(\mathbf{A})$ is the spectral norm of the matrix $\mathbf{A}$.

Therefore, to show condition C1, it suffices to show that

$$\max_{l=1,\cdots,m}\lambda_{\max}\left[\frac{1}{N}\mathbf{\Psi}^T\mathbf{\Psi} - \frac{1}{n}\mathbf{\Psi}_l^T\mathbf{\Psi}_l\right] = o_{\mathbb{P}_X}(1). \tag{A.13}$$

Using Lemma 2 of Bach (2013), we have that

$$\mathbb{P}_I\left(\lambda_{\max}\left[\frac{1}{N}\mathbf{\Psi}^T\mathbf{\Psi} - \frac{1}{n}\mathbf{\Psi}_l^T\mathbf{\Psi}_l\right] > t\right) \leq r\exp\left(\frac{-nt^2/2}{\lambda_{\max}\left[\frac{1}{N}\mathbf{\Psi}^T\mathbf{\Psi}\right](R^2 + t/3)}\right), \tag{A.14}$$

where $\mathbb{P}_I$ is the probability measure corresponding to the partition of the data, $r = \text{rank}(\mathbf{\Psi}) = \text{rank}(\mathbf{K})$ and R is the upperbound of L2-norm of all rows of $\mathbf{\Psi}$. In our case, L2-norm of all rows of $\mathbf{\Psi}$ the diagonal elements of matrix

$$\mathbf{\Psi}\mathbf{\Psi}^T = \mathbf{\Phi}\left(\lambda \mathbf{I} + \frac{1}{N}\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T = N\mathbf{K}(\mathbf{K} + N\lambda\mathbf{I})^{-1},$$

where the last equality follows from the Sherman–Morrison formula. Then, by the definition of d_λ in (20), we have that $R^2 \leq d_\lambda$. In addition, note that

$$\lambda_{\max}\left(\frac{1}{N}\mathbf{\Psi}^T\mathbf{\Psi}\right) = \lambda_{\max}\left(\frac{1}{N}\mathbf{\Phi}\left(\lambda \mathbf{I} + \frac{1}{N}\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\right) \leq 1,$$

which implies that inequality (A.14) can be further simplified as

$$\mathbb{P}_I\left(\lambda_{\max}\left[\frac{1}{N}\mathbf{\Psi}^T\mathbf{\Psi}-\frac{1}{n}\mathbf{\Psi}_l^T\mathbf{\Psi}_l\right]>t\right)\leq r\exp\left(\frac{-nt^2/2}{d_\lambda+t/3}\right),$$

which further leads to that

$$\mathbb{P}_I\left(\max_{l=1,\cdots,m}\lambda_{\max}\left[\frac{1}{N}\mathbf{\Psi}^T\mathbf{\Psi}-\frac{1}{n}\mathbf{\Psi}_l^T\mathbf{\Psi}_l\right]>t\right)\leq mr\exp\left(\frac{-nt^2/2}{d_\lambda+t/3}\right)\to^{\mathbb{P}_X}0,$$

for any $0<t<3d_\lambda$ under condition C1', which completes the proof of (A.13). □

Proof of Lemma 3. We first consider $Q_2(\lambda|\mathbf{X})$ in (24). Define the function class

$$\mathcal{F}_n=\left\{f(x):\|f\|_{\sup}\leq C_1V_n, J_1(f)\leq C_2V_n^2H_n^2\right\},$$

where V_n and H_n are as defined in Conditions C4'(b)-(c) and C_1, C_2 are some constants. Applying Lemma A.3 to the function class $\mathcal{F}_n$ with $\epsilon_n=\epsilon$ and $\gamma_n=\sqrt{\frac{c_0(1+H_n)V_n}{n}}$, which satisfy (A.12) under Conditions C4'(b)-(c), we have that

$$P\left(\sup_{f\in\mathcal{V}_n}\frac{|\mathbb{P}_nf-\mathbb{P}f|}{\mathbb{P}_nf+\mathbb{P}f+\gamma_n}>C\epsilon\right)\leq\exp(-n\epsilon^2\gamma_n/2). \tag{A.15}$$

Let $v_k(x)=\mathrm{Var}_\varepsilon\left\{\widehat{f}_k(x)\right\}$, $k=1,\ldots,m$. It is straightforward to show that the first derivative of $v_k(x)$ are bounded as follows

$$|v_k'(x)|=2\left|\mathrm{Cov}_\varepsilon\left\{\widehat{f}_k(x),\widehat{f}_k'(x)\right\}\right|\leq 2\sqrt{v_k(x)}\sqrt{\mathrm{Var}_\varepsilon\{\widehat{f}_k'(x)\}},$$

which further implies that

$$\begin{aligned}J_1(v_k)&=\int_{\mathcal{X}}\{v_k'(x)\}^2\,d\mathbb{P}_X(x)\leq 4\|v_k\|_{\sup}\int_{\mathcal{X}}\mathrm{Var}_\varepsilon\{\widehat{f}_k'(x)\}\,d\mathbb{P}_X(x)\\&\leq 4\|v_k\|_{\sup}^2\frac{\int_{\mathcal{X}}\mathrm{Var}_\varepsilon\{\widehat{f}_k'(x)\}\,d\mathbb{P}_X(x)}{\int_{\mathcal{X}}v_k(x)\,d\mathbb{P}_X(x)}=O_{\mathbb{P}_X}(V_n^2H_n^2).\end{aligned}$$

Therefore, under conditions C4'(a)-(b), we have that

$$v_1(x),\ldots,v_m(x)\in\mathcal{F}_n \text{ in probability measure } \mathbb{P}_X.$$

For simplicity, from now on, we use Q for $Q(\lambda|\mathbf{X})$ in (22) and Q_j for $Q_j(\lambda|\mathbf{X})$, $j=1,2$, in (23) and (24) whenever there is no ambiguity. Using the facts that $Q=\frac{1}{m^2}\sum_{k=1}^m\mathbb{P}v_k$ and $Q_2=\frac{1}{m^2}\sum_{k=1}^m\mathbb{P}_{n_k}v_k$, a direct application of (A.15) gives that

$$\begin{aligned}P\left(\frac{|Q_2-Q|}{Q_2+Q+\frac{1}{m}\gamma_n}>C\epsilon\right)&\leq P\left(\frac{\frac{1}{m}\sum_{k=1}^m|\mathbb{P}_{n_k}v_k-\mathbb{P}_{n_k}v_k|}{\frac{1}{m}\sum_{k=1}^m(\mathbb{P}_{n_k}v_k+\mathbb{P}_{n_k}v_k)+\gamma_n}>C\epsilon\right)\\&\leq P\left(\max_{1\leq k\leq m}\left(\frac{|\mathbb{P}_{n_k}v_k-\mathbb{P}_{n_k}v_k|}{\mathbb{P}_{n_k}v_k+\mathbb{P}_{n_k}v_k+\gamma_n}\right)>C\epsilon\right)\\&\leq m\exp(-n\epsilon^2\gamma_n/2)\to 0,\end{aligned}$$

where the last step follows from condition C4'(c). In addition, by conditions C4'(b)-(c), we have that $\frac{\gamma_n}{mQ} = \sqrt{\frac{c_0(1+H_n)V_n}{mNQ^2}} = O_{\mathbb{P}_X}(1)$. Hence we conclude that

$$Q_2(\lambda|\mathbf{X}) = Q(\lambda|\mathbf{X}) + o_{\mathbb{P}_X} Q\{(\lambda|\mathbf{X})\}. \tag{A.16}$$

Now we turn to the quantity $Q_1(\lambda|\mathbf{X})$. Define another function class

$$\bar{\mathcal{F}}_n = \left\{ f(x) : \|f\|_{\sup} \leq C_1 \frac{V_n}{m}, J_1(f) \leq C_2 \frac{V_n^2 H_n^2}{m^2} \right\},$$

where V_n and H_n are as defined in Conditions C4'(b)-(c) and C_1, C_2 are some constants. By applying Lemma A.3 to the function class $\bar{\mathcal{F}}_n$ with $\epsilon_n = \epsilon$ and $\gamma_N = \sqrt{\frac{c_0(1+H_n)V_n}{mN}}$, which satisfy (A.12) under Conditions C4'(b)-(c), we have that

$$P\left(\sup_{f \in \mathcal{V}_N} \frac{|\mathbb{P}_N f - \mathbb{P} f|}{\mathbb{P}_N f + \mathbb{P} f + \gamma_N} > C\epsilon \right) \leq \exp(-N\epsilon^2 \gamma_N / 2). \tag{A.17}$$

Define another function

$$\bar{v}(x) = \mathrm{Var}_\varepsilon\{\bar{f}(x)\} = \frac{1}{m^2} \sum_{k=1}^m v_k(x),$$

whose derivative is bounded as

$$\begin{aligned} |\bar{v}'(x)| &= 2 \left| \mathrm{Cov}_\varepsilon \left\{ \bar{f}(x), \bar{f}'(x) \right\} \right| \leq \frac{2}{m} \sqrt{\mathrm{Var}_\varepsilon\{\bar{f}(x)\}} \sqrt{\mathrm{Var}_\varepsilon\{\bar{f}'(x)\}} \\ &\leq \frac{2}{m} \sqrt{\frac{1}{m} \sum_{k=1}^m v_k(x)} \sqrt{\frac{1}{m} \sum_{k=1}^m \mathrm{Var}_\varepsilon\{\widehat{f}_k'(x)\}}. \end{aligned}$$

From the above two equations/inequalities, under conditions C4'(b)-(c), one has that

$$\|\bar{v}\|_{\sup} \leq \frac{1}{m^2} \sum_{k=1}^m \|v_k\|_{\sup} \frac{1}{m} O_{\mathbb{P}_X}(V_n),$$

and that

$$\begin{aligned} J_1(\bar{v}) &= \int_{\mathcal{X}} \{\bar{v}_k'(x)\}^2 \, d\mathbb{P}_X(x) \bar{v} \leq \frac{4}{m^2} \int_{\mathcal{X}} \left\{ \frac{1}{m} \sum_{k=1}^m v_k(x) \right\}^2 \frac{\frac{1}{m} \sum_{k=1}^m \mathrm{Var}_\varepsilon\{\widehat{f}_k'(x)\}}{\frac{1}{m} \sum_{k=1}^m v_k(x)} \, d\mathbb{P}_X(x) \\ &\leq \frac{4}{m^2} \left\{ \max_{1 \leq k \leq m} \|v_k\|_{\sup} \right\}^2 \int_{\mathcal{X}} \max_{1 \leq k \leq m} \frac{\mathrm{Var}_\varepsilon\{\widehat{f}_k'(x)\}}{v_k(x)} \, d\mathbb{P}_X(x) = \frac{1}{m^2} O_{\mathbb{P}_X}(V_n^2 H_n^2) \end{aligned}$$

Therefore, under conditions C4'(a)-(b), we have that

$$\bar{v}(x) \in \bar{\mathcal{F}}_n \text{ in probability measure } \mathbb{P}_X.$$

Using the facts that $Q = \mathbb{P}\bar{v}$ and $Q_1 = \mathbb{P}_N\bar{v}$, a direct application of (A.17) gives that

$$P\left(\frac{|Q_1 - Q|}{Q_1 + Q + \gamma_N} > C\epsilon\right) = P\left(\sup_{\bar{v}\in\bar{\mathcal{V}}_N} \frac{|\mathbb{P}_N\bar{v} - \mathbb{P}\bar{v}|}{\mathbb{P}_N\bar{v} + \mathbb{P}\bar{v} + \gamma_N} > C\epsilon\right) \leq \exp(-N\epsilon^2\gamma_N/2) \to 0,$$

where the last step follows from condition C4'(c). Furthermore, by conditions C4'(b)-(c), we have that $\frac{\gamma_N}{Q} = \sqrt{\frac{c_0(1+H_n)V_n}{mNQ^2}} = O_{\mathbb{P}_X}(1)$. Hence we conclude that

$$Q_1(\lambda|\mathbf{X}) = Q(\lambda|\mathbf{X}) + o_{\mathbb{P}_X}\{Q(\lambda|\mathbf{X})\}. \tag{A.18}$$

Combining equations (A.16)–(A.18), we have that

$$\frac{\frac{1}{Nm}\sum_{k=1}^{m}\mathrm{tr}(\mathbf{A}_{kk}^2)}{\frac{\mathrm{tr}(\bar{\mathbf{A}}_m^T\bar{\mathbf{A}}_m)}{N}} = \frac{Q_1(\lambda|\mathbf{X})}{Q_2(\lambda|\mathbf{X})} = O_{\mathbb{P}_X}(1). \tag{A.19}$$

By the definition of $\bar{\mathbf{A}}_m$, it is straightforward to show that

$$\frac{\{\frac{1}{N}\mathrm{tr}(\bar{\mathbf{A}}_m)\}^2}{\frac{1}{Nm}\sum_{k=1}^{m}\mathrm{tr}(\mathbf{A}_{kk}^2)} = \frac{1}{N}\frac{\{\frac{1}{m}\sum_{k=1}^{m}\mathrm{tr}(\mathbf{A}_{kk})\}^2}{\frac{1}{m}\sum_{k=1}^{m}\mathrm{tr}(\mathbf{A}_{kk}^2)} \leq \frac{1}{N}\frac{1}{m}\sum_{k=1}^{m}\frac{\{\mathrm{tr}(\mathbf{A}_{kk})\}^2}{\mathrm{tr}(\mathbf{A}_{kk}^2)} = \frac{1}{m}\sum_{k=1}^{m}\frac{\{N^{-1}\mathrm{tr}(\mathbf{A}_{kk})\}^2}{N^{-1}\mathrm{tr}(\mathbf{A}_{kk}^2)},$$

where the second last inequality follows from Cauchy-Schwartz inequality. Combining the above inequality and (A.19), under condition C4'(a), we finally have that

$$\frac{\{\frac{1}{N}\mathrm{tr}(\bar{\mathbf{A}}_m)\}^2}{\{\frac{1}{N}\mathrm{tr}(\bar{\mathbf{A}}_m^T\bar{\mathbf{A}}_m)\}} = \frac{\{\frac{1}{N}\mathrm{tr}(\bar{\mathbf{A}}_m)\}^2}{\frac{1}{Nm}\sum_{k=1}^{m}\mathrm{tr}(\mathbf{A}_{kk}^2)}\frac{\frac{1}{Nm}\sum_{k=1}^{m}\mathrm{tr}(\mathbf{A}_{kk}^2)}{\{\frac{1}{N}\mathrm{tr}(\bar{\mathbf{A}}_m^T\bar{\mathbf{A}}_m)\}} = o_{\mathbb{P}_X}(1),$$

which completes the proof. □