Optimal tuning for divide-and-conquer kernel ridge regression with massive data

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

October 26, 2016

Abstract

We propose a first data-driven tuning procedure for divide-and-conquer kernel ridge regression (Zhang et al., 2015). While the proposed criterion is computationally scalable for massive data sets, it is also shown to be asymptotically optimal under mild conditions. The effectiveness of our method is illustrated by extensive simulations and an application to Million Song Dataset.

Some key words: Distributed GCV, divide-and-conquer, kernel ridge regression, optimal tuning.

Short title: Optimal tuning for divide-and-conquer kernel ridge regression.

[1]Department of Mathematical Sciences, Binghamton University, The State University of New York, Binghamton, NY 13902, USA.
E-mail: gang@math.binghamton.edu

[1]Department of Mathematical Sciences, Binghamton University, The State University of New York, Binghamton, NY 13902, USA.
E-mail: zshang@math.binghamton.edu

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

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 With a large to massive sample size, many sophisticated statistical tools are no longer applicable simply due to formidable computational costs and/or memory requirements.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. For example, divide-and-conquer kernel ridge regression has recently been developed in the environment of parallel and distributed computation (Zhang et al., 2015). In this paper, we propose a first data-driven tuning method for the above nonparametric estimation with statistical guarantee.

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

$$y_i = f_0(x_i) + \varepsilon_i, \quad i = 1, \ldots, N, \tag{1}$$

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 & Cristianini, 2004) which essentially aims at finding a projection of $f_0(\cdot)$ into a reproducing kernel Hilbert space (RKHS), denoted as $\mathcal{H}$, with a norm $\|\cdot\|_{\mathcal{H}}$. Specifically, the kernel ridge regression estimator can be obtained as follows

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

where $\lambda \to 0$ 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). When the sample size N is large, such requirements can be prohibitive even for an advanced computer. To overcome this computation bottleneck, Zhang et al. (2015) proposed the following "divide-and-conquer" algorithm: (i) Randomly divide the entire sample $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ to m disjoint "smaller" subsets, denoted by $S_1, \ldots, S_m$; (ii) For each subset S_k, find $\widehat{f}_k = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n_k} \sum_{i \in S_k} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}} \right\}$, where n_k is the size of S_k; (iii) The final nonparametric estimator is given by

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

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 could 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 uses 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. If we naively apply traditional tuning methods, e.g., Mallow's CP (Mallows, 1973), Generalized cross-validation (Craven & Wahba, 1979, GCV) and Generalized approximated cross-validation (Xiang & Wahba, 1996), 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 Zhang et al. (2015), 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 $k = 1, \ldots, m$. However, how to obtain such an optimal λ empirically remains unclear.

In this paper, we define a new data-driven criterion named "distributed generalized cross-validation" (dGCV) to choose tuning parameters for kernel ridge regression 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 & Wahba, 1979) 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 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 & 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

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

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 $K(x,z) = \sum_{j=1}^{\infty} \mu_j \psi_j(x) \psi_j(z)$, 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 scalars x, z, the polynomial kernel $K(x, z) = (1 + xz)^r$ 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 Gaussian kernel $K(x, z) = \exp(-\|x - z\|_2^2/\phi^2)$, 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.

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

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

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

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

where $\boldsymbol{Y} = (y_1, \ldots, y_N)^T$ and $\mathbf{K} = [K(x_i, x_j)]_{i,j=1,\ldots,N}$.

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

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

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

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

$$L_k(\lambda|\boldsymbol{x}_k) = \frac{1}{n_k}\sum_{i=1}^{n_k} w_{ki}\left\{\widehat{f}_k(x_{k,i}) - f_0(x_{k,i})\right\}^2, \tag{7}$$

where $w_{ki} \geq 0$ is some weight assigned to each observation $(y_{k,i}, x_{k,i})$ and satisfies $\sum_{i=1}^{n_k} w_{ki} = n_k$. The introduction of weights in (7) 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 $L_k(\lambda|\boldsymbol{x}_k)$. When only focused on a single sub-data set, such a "locally-optimal" choice of tuning parameter λ has been well studied in (Craven & Wahba, 1979; Li, 1986; Gu, 2002; Wood, 2004; Gu& Ma, 2005; Xu & Huang, 2012), among which the most popular method remains to be the Generalized Cross-Validation (Craven & Wahba, 1979).

Using the function estimator $\widehat{f}_k(x)$, the predicted values for the vector $\boldsymbol{y}_k$ can be written as $\widehat{\boldsymbol{y}}_k = \mathbf{A}_{kk}(\lambda)\boldsymbol{y}_k$, 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

$$\mathrm{GCV}_k(\lambda) = \frac{n_k^{-1}(\widehat{\boldsymbol{y}}_k - \boldsymbol{y}_k)^T \mathbf{W}_k (\widehat{\boldsymbol{y}}_k - \boldsymbol{y}_k)}{\{1 + n_k^{-1}\mathrm{tr}\{\mathbf{A}_{kk}(\lambda)\mathbf{W}_k\}\}^2}, \tag{8}$$

where $\mathbf{W}_k = \mathrm{diag}\{w_{k1}, \ldots, w_{kn_k}\}$, $k = 1, \ldots, m$. It is well known that $\mathrm{GCV}_k(\lambda)$ enjoys nice asymptotic properties. For example, under mild conditions, Gu (2002) showed that, as $n_k \to \infty$,

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

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

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 $R_k(\lambda|\boldsymbol{x}_k) = \mathbb{E}_\varepsilon\left\{L_k(\lambda|\boldsymbol{x}_k)\right\}$ is of the form

$$R_k(\lambda|\boldsymbol{x}_k) \quad = \quad \frac{1}{n_k}\sum_{i=1}^{n_k} w_i \mathrm{Var}_\varepsilon\left\{\widehat{f}_k(x_{k,i})\right\} + \frac{1}{n_k}\sum_{i=1}^{n_k} w_i \left\{\mathbb{E}_\varepsilon \widehat{f}_k(x_{k,i}) - f_0(x_{k,i})\right\}^2, \tag{9}$$

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

$$\bar{L}(\lambda|\boldsymbol{X}) = \frac{1}{N}\sum_{i=1}^{N} w_i\{\bar{f}(x_i) - f_0(x_i)\}^2, \tag{10}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$ 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 $\bar{R}(\lambda|\boldsymbol{X}) = \mathbb{E}_\varepsilon\{\bar{L}(\lambda|\boldsymbol{X})\}$ has the following form

$$\bar{R}(\lambda|\boldsymbol{X}) = \frac{1}{m^2 N}\sum_{j=1}^{m}\sum_{i=1}^{N} w_i \mathrm{Var}_\varepsilon\left\{\widehat{f}_k(x_i)\right\} + \frac{1}{N}\sum_{i=1}^{N} w_i \left[\frac{1}{m}\sum_{j=1}^{m}\left\{\mathbb{E}_\varepsilon \widehat{f}_k(x_i) - f_0(x_i)\right\}\right]^2. \tag{11}$$

The form of (9) illustrates that, roughly speaking, a "locally optimal" choice of λ (that minimizes (7)) 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 (10), 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 (11). 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 $\boldsymbol{y}_l$ in the subset S_l using $\widehat{f}_k$ take the form $\widehat{\boldsymbol{y}}_{kl} = \mathbf{A}_{kl}\boldsymbol{y}_k$, 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 $\boldsymbol{Y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_m^T)^T$. Then the predicted value of $\boldsymbol{Y}$ using the averaged estimator $\bar{f}$ is of the form

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

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

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

Furthermore, the global conditional risk function (11) can be conveniently re-written as

$$\bar{R}(\lambda|\boldsymbol{X}) = \frac{1}{N}\boldsymbol{F}^T\{\mathbf{I} - \bar{\mathbf{A}}_m(\lambda)\}^T\mathbf{W}\{\mathbf{I} - \bar{\mathbf{A}}_m(\lambda)\}\boldsymbol{F} + \frac{\sigma^2}{N}\mathrm{tr}\left\{\bar{\mathbf{A}}_m^T(\lambda)\mathbf{W}\bar{\mathbf{A}}_m(\lambda)\right\}, \qquad (12)$$

where $\boldsymbol{F} = (\boldsymbol{f}_1^T, \ldots, \boldsymbol{f}_m^T)^T$ 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 (2002), we can define an unbiased estimator of $\bar{R}(\lambda|\boldsymbol{X}) + \sigma^2$ as follows

$$\bar{U}(\lambda|\boldsymbol{X}) = \frac{1}{N}\boldsymbol{Y}^T\{\mathbf{I} - \bar{\mathbf{A}}_m(\lambda)\}^T\mathbf{W}\{\mathbf{I} - \bar{\mathbf{A}}_m(\lambda)\}\boldsymbol{Y} + \frac{2\sigma^2}{N}\mathrm{tr}\left\{\bar{\mathbf{A}}_m(\lambda)\mathbf{W}\right\}. \qquad (13)$$

It is straightforward to show that $\mathbb{E}_\varepsilon\{\bar{U}(\lambda|\boldsymbol{X})\} = \bar{R}(\lambda|\boldsymbol{X}) + \sigma^2$. The above $\bar{U}(\lambda|\boldsymbol{X})$ can be viewed as an extension of the Mallow's CP (Mallows, 1973) to the divide-and-conquer scenario.

Similar to Gu (2002); Xu & Huang (2012), the Lemma 1 in Section 4 states that under some mild conditions, minimizing $\bar{U}(\lambda|\boldsymbol{X})$ and $\bar{L}(\lambda|\boldsymbol{X})$ with respect to λ is asymptotically equivalent. In this sense, the λ chosen by minimizing $\bar{U}(\lambda|\boldsymbol{X})$ is therefore "globally optimal." However, a major drawback of $\bar{U}(\lambda|\boldsymbol{X})$ 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

$$\mathrm{dGCV}(\lambda|\boldsymbol{X}) = \frac{\frac{1}{N}\sum_{i=1}^N w_i\left\{y_i - \bar{f}(x_i)\right\}^2}{\left[1 - \frac{1}{Nm}\sum_{k=1}^m \mathrm{tr}\{\mathbf{A}_{kk}(\lambda)\mathbf{W}_k\}\right]^2}, \qquad (14)$$

where $\mathbf{W}_k = \{w_i\}_{i\in S_k}$. 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 $\bar{U}(\lambda|\boldsymbol{X})$ defined in (13) essentially can be viewed as the first order Taylor expansion of the $\mathrm{dGCV}(\lambda|\boldsymbol{X})$. However, in the definition of $\mathrm{dGCV}(\lambda|\boldsymbol{X})$, 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 $\mathrm{dGCV}(\lambda|\boldsymbol{X})$ and the "golden criterion" $\bar{L}(\lambda|\boldsymbol{X})$ with respect to

λ are also asymptotically equivalent.

3.4 Computational Complexity of dGCV

The computation of dGCV($\lambda|\boldsymbol{X}$) in (14) 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 dGCV($\lambda|\boldsymbol{X}$) is of the order $O(N^3/m^2 + NN_w)$. In most applications, the number of folds m generally cannot exceed $\sqrt{N}$ for $\bar{f}$ to reach the optimal convergence rate (Zhang et al., 2015). In such cases, 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 dGCV($\lambda|\boldsymbol{X}$). 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 dGCV($\lambda|\boldsymbol{X}$) becomes $O(NN_w)$. In this case, we may want to only choose K 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 $w_i = N/(\sum_{k=K}^{K} n_k)$ if $i \in \cup_{k=1}^{K} S_k$ and $w_i = 0$ otherwise. Under this setting, the dGCV($\lambda|\boldsymbol{X}$) in (14) becomes

$$\mathrm{dGCV}^*(\lambda|\boldsymbol{X}) = \frac{\frac{1}{N_K}\sum_{i\in\cup_{k=1}^{K} S_k}\left\{y_i - \bar{f}(x_i)\right\}^2}{\left[1 - \frac{1}{mN_K}\sum_{k=1}^{K}\mathrm{tr}\{\mathbf{A}_{kk}(\lambda)\}\right]^2}, \quad N_K = n_1 + \cdots + n_K. \tag{15}$$

Using (15) instead of (14), we only need to evaluate $\bar{f}(x_i)$ for x_i's in K subsets and the computation time is reduced to $O(N^2K/m)$. We applied (15) to the Million Song Data set considered in Section 6, which yields good results in both prediction and computation time.

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

Remark 1. *We want to mention that dGCV($\lambda|\boldsymbol{X}$) can also be used to choose other tuning parameters in the kernel function. For example, if the Gaussian kernel $K(x,z) = \exp(\|x-z\|_2^2/\phi)$ is used, dGCV is also a function of the bandwidth parameter ϕ, and thus can be used to choose the optimal ϕ as well. In section 6, a large reduction of prediction error was achieved for the Million Song Dataset by using such a choice of ϕ.*

4 Asymptotic Properties

In this section, we will show that the proposed dGCV criterion in (14) 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 $\text{tr}(\mathbf{A})$ 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]$\frac{1}{m}\sum_{l=1}^m \lambda_{\max}\left\{(\mathbf{K}_{ll}+\lambda\mathbf{I}_l)^{-2}\left(\frac{1}{m}\sum_{k=1}^m \mathbf{K}_{kl}^T\mathbf{K}_{kl}\right)\right\} = O_{\mathbb{P}_X}(1)$;

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

[C3] (a) $\max_{1\le i\le N} w_i \le W$ for some $W > 0$; (b) $\frac{1}{Nm}\sum_{k=1}^{m} \text{tr}\{\mathbf{A}_{kk}(\lambda)\} = o_{\mathbb{P}_X}(1)$.

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

Intuitively, condition C1 requires some similarities among sub-data sets. If all $\mathbf{K}_{kl}$'s are similar to the the matrix $\mathbf{K}_{ll}$, we can expect $\lambda_{\max}\left\{(\mathbf{K}_{ll}+\lambda\mathbf{I}_l)^{-2}\left(\frac{1}{m}\sum_{k=1}^{m}\mathbf{K}_{kl}^T\mathbf{K}_{kl}\right)\right\} \le 1$, in which case C1 holds. Condition C2 is a widely used condition to ensure the optimality of the GCV to hold, for example, see Craven & Wahba (1979); Li (1986); Gu& Ma (2005); Xu & 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 $\bar{R}(\lambda|\boldsymbol{X})$ 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). It turns out that, under conditions C1-C2, $\bar{U}(\lambda|\boldsymbol{X})$ defined in (13) is "globally optimal."

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

$$\bar{U}(\lambda|\boldsymbol{X}) - \bar{L}(\lambda|\boldsymbol{X}) - \frac{1}{N}\boldsymbol{\varepsilon}^T\mathbf{W}\boldsymbol{\varepsilon} = o_{\mathbb{P}_{\varepsilon,X}}\{\bar{L}(\lambda|\boldsymbol{X})\}. \tag{16}$$

The proof is given in the Appendix.

Lemma 1 states that when σ^2 is known, minimizing $\bar{U}(\lambda|\boldsymbol{X})$ with respect to λ is asymptotically equivalent to minimizing the empirical true loss function $\bar{L}(\lambda|\boldsymbol{X})$. 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*

$$\mathrm{dGCV}(\lambda|\mathbf{X}) - \bar{L}(\lambda|\mathbf{X}) - \frac{1}{N}\boldsymbol{\varepsilon}^T\mathbf{W}\boldsymbol{\varepsilon} = o_{\mathbb{P}_{\varepsilon,X}}\{(\bar{L}(\lambda|\mathbf{x})\}. \tag{17}$$

The proof is given in the Appendix.

Similar to Lemma 1, Theorem 1 shows that minimizing $\mathrm{dGCV}(\lambda|\boldsymbol{X})$ amounts to minimizing the true conditional loss function $\bar{L}(\lambda|\boldsymbol{X})$, although additional conditions C3-C4 are needed. Condition C3 is pretty mild in that it essentially requires that sufficient number of w_i's are nonzero and the effective number of parameters to be negligible compared to the sample size, which is typically true for non-parametric function estimators in most settings of interests. In addition, C3 becomes trivial when $m \to \infty$ because by definition we have that $\mathrm{tr}\{\mathbf{A}_{kk}(\lambda)\} \leq n_k$, $k = 1, \ldots, m$. Condition C4 will be discussed in more details in the next subsection.

4.2 Validation of Condition C4

We shall only consider uniform weights with $w_1 = \cdots = w_N = 1$ and equal sample sizes $n_1 = \cdots = n_m = n$ in this subsection. 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 & Wahba, 1979; Li, 1986; Gu& Ma, 2005; Xu & 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 interests 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

$$Q(\lambda|\boldsymbol{X}) = \int_{\mathcal{X}} \mathrm{Var}_{\varepsilon}\{\bar{f}(x)\}^2 \, d\mathbb{P}_X(x) = \frac{1}{m^2}\sum_{k=1}^{m}\int_{\mathcal{X}} \mathrm{Var}_{\varepsilon}\{\widehat{f}_k(x)\} \, d\mathbb{P}_X(x). \tag{18}$$

Let $\mathbb{P}_{X,N}$ be the empirical measure based on sample $\{X_1, \dots, X_N\}$, 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

$$Q_1(\lambda|\boldsymbol{X}) = \sigma^2 \frac{\mathrm{tr}\{\bar{\mathbf{A}}_m^T(\lambda)\bar{\mathbf{A}}_m(\lambda)\}}{N} = \int_{\mathcal{X}} \mathrm{Var}_{\varepsilon}\left\{\bar{f}(x)\right\}^2 \, d\mathbb{P}_{X,N}(x), \tag{19}$$

$$Q_2(\lambda|\boldsymbol{X}) = \sigma^2 \frac{1}{Nm}\sum_{k=1}^{m} \mathrm{tr}\{\mathbf{A}_{kk}^2(\lambda)\} = \frac{1}{m^2}\sum_{k=1}^{m}\int_{\mathcal{X}} \mathrm{Var}_{\varepsilon}\{\widehat{f}_k(x)\} \, d\mathbb{P}_{X,n_k}(x). \tag{20}$$

Intuitively, $Q_1(\lambda|\boldsymbol{X})$ and $Q_2(\lambda|\boldsymbol{X})$ are two empirical versions of $Q(\lambda|\boldsymbol{X})$ and should be close to each other. The formal proof utilizes the uniform ratio limit theorems for empirical processes (Pollard, 1995) to show $Q_1(\lambda|\boldsymbol{X})/Q(\lambda|\boldsymbol{X}) = 1 + o_{\mathbb{P}_X}(1)$ and $Q_2(\lambda|\boldsymbol{X})/Q(\lambda|\boldsymbol{X}) = 1 + o_{\mathbb{P}_X}(1)$, 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 $\mathcal{F}_0 = \{f : \|f\|_{\sup} \leq 1, J_1(f) = \int_{\mathcal{X}} \{f'(x)\}^2 \, d\mathbb{P}_X(x) \leq 1\}$, we have that $\mathcal{N}(\epsilon, \|\cdot\|_{\mathbb{P}_{X,n}}, \mathcal{F}_0) \leq \exp(C_0/\epsilon)$ for some constant $C_0 > 0$ with probability approaching one as $n \to \infty$.

Lemma 2. *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 supreme 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 (9) 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 are 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, for example van der 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 $Q_1(\lambda|\boldsymbol{X})$ and $Q_2(\lambda|\boldsymbol{X})$ to $Q(\lambda|\boldsymbol{X})$. And it can be relaxed similarly as given in Theorem 2.1 of Pollard (1986).

5 Simulation studies

In this section, we conduct a simulation study to illustrate the effectiveness of dGCV(λ) in choosing the optimal λ for the divide-and-conquer function estimator. The data were simulated from the model $y = f_0(x) + \varepsilon$, where $f_0(x) = 2|x - 1/2|$ for $x \in [0, 1]$ and $\varepsilon \sim N(0, 0.5^2)$. 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). The true function $f_0(x)$ belongs to the Sobolev space of Lipschitz functions on $[0, 1]$, hence we used the reproducing kernel $K(x, z) = 1 + \min(x, z)$ and the associated norm $\|f\|^2_{\mathcal{H}} = f^2(0) + \int_0^1 \{f'(x)\}^2\, dx$.

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 $[-12, 1]$. Three approaches were used for selection of λ: (i) the distributed GCV (dGCV) proposed in (14); (ii) the naive GCV applied to each sub-dataset (nGCV) which applies the GCV defined in (8) to choose the best λ for each individual $\widehat{f}_k$ and then average them using (3); and (iii) the true conditional loss function (TrueLoss) $\bar{L}(\lambda|\boldsymbol{X})$ defined in (10). For all three approaches, we set the weights $w_i = 1$ for all $i = 1, \ldots, N$. The last approach is not practically feasible since it requires the knowledge of the true function f_0. Rather, it is served as the "golden criterion" to show the effectiveness of the other two approaches. 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 dGCV(λ) for $N = 2^i$, $i = 8, 9, 10, 11, 12$ and $m = 1, 2, 4, 8, 16, 32$. All simulation runs were carried out in the software R 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 (when $m = 1$).

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 (10), of the function estimators obtained by minimizing dGCV(λ) versus minimizing the unattainable "golden criterion" (10) 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."

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 distributed GCV and the naive GCV 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 naive GCV 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 consist 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 $O(n^{-2/3})$ with $n = N/m$ whereas the globally optimal rate for λ is of the order $O(N^{-2/3})$.

The y-axis of Figure 1(d)-(e) 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 (10) 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.

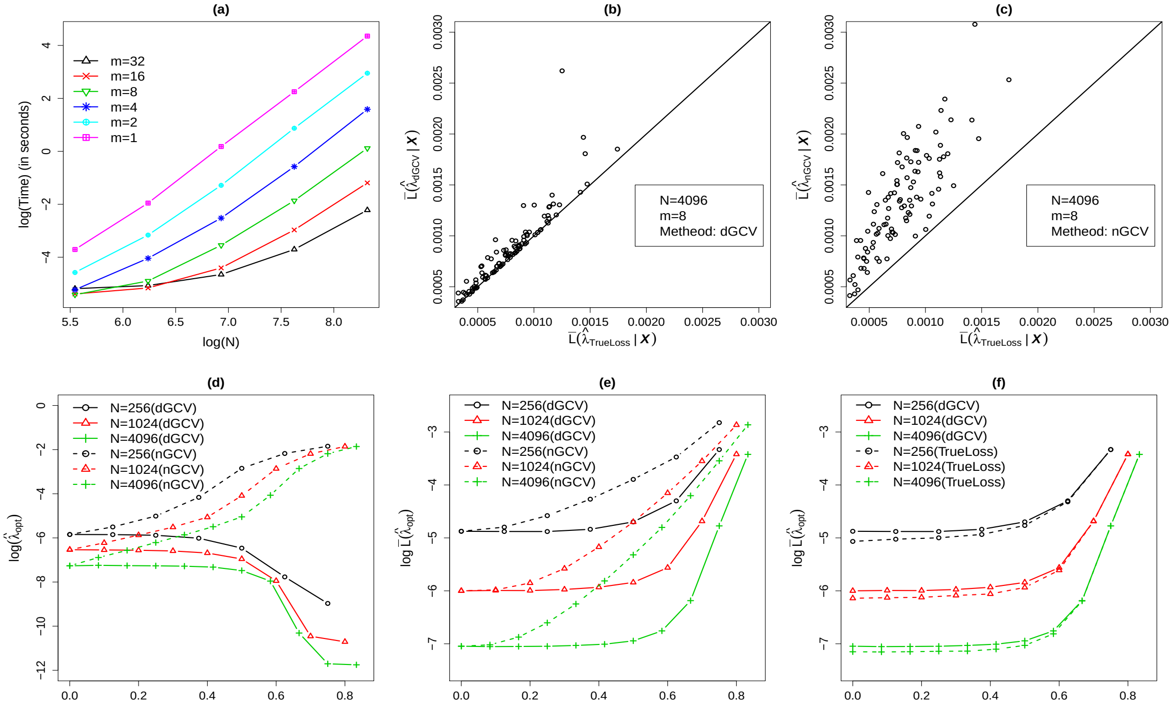

Figure 1: (a) the logarithm of computational time (in seconds) v.s. $\log(N)$; (b)-(c): scatter plots of true empirical losses of function estimators; (d) the logarithm of averages of selected λ v.s. $\log(m)/\log(N)$; (e)-(f): the logarithm of true empirical losses v.s. $\log(m)/\log(N)$. Note that 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.

6 A real data example

In this section, we applied the dGCV* tuning method to the Million Song Dataset, which consist 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. The goal is to use the timbre information of the song to predict the year when the song was released using the kernel ridge regression. 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 kernel ridge regression. Seven partitions $m \in \{32, 38, 48, 64, 96, 128, 256\}$ were used for the divide-and-conquer kernel ridge regression. 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 (15) with $K = \lceil m/10 \rceil$, where $\lceil a \rceil$ is the smallest integer that is greater than a. Note that in this case, dGCV$^*(\lambda|\boldsymbol{X})$ is also a function of ϕ. The experiment was conducted in Matlab using a Windows computer with 16GB of memory and a single-threaded 3.5Ghz CPU. For each (λ, ϕ) pair, the computation time for (15) are $1,058$s ($m = 32$), 840s ($m = 38$), 577s ($m = 48$), 476s ($m = 96$), 453s ($m = 128$) and 457s ($m = 256$), which are reasonable for a data set with almost half-million observations.

The grid search gave the optimal choice of $\lambda = 0.5/N$ and $\phi = 3$ for most of case scenarios. From Figure 2(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.

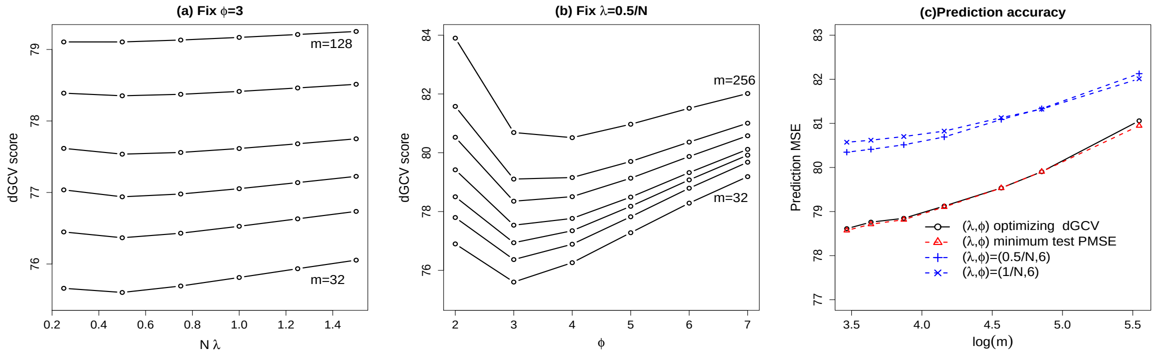

Figure 2: (a) dGCV score v.s. $N\lambda$ with $m = 32$ (the bottommost) to $m = 128$ (the uppermost); (b) dGCV score v.s. ϕ with $m = 32$ (the bottommost) to $m = 256$ (the uppermost); (c) The prediction mean squared errors on the testing samples v.s. $\log(m)$.

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 2(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.

Acknowledgment

Guang Cheng's research is partially supported by NSF CAREER Award DMS-1151692, DMS-1418042 and Office of Naval Research (ONR N00014-15-1-2331).

References

Bertin-Mahieux, T., Ellis, D.P., Whitman, B. and Lamere, P. (2011). The million song dataset. *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR2011).*

Craven, P. and Wahba, G. (1979). Smoothing Noisy Data with Spline Functions. *Numer. Math.*, **31**, 377–403.

Gu, C. (2002). *Smoothing Spline ANOVA Models.* Springer: New York.

Gu, C. and Ma, P. (2005). Optimal Smoothing in Nonparametric Mixed-Effect Models. *Ann. Statist.*, **33**, 1357–79.

Mallows, C.L. (1973). Some Comments on CP. *Technometrics*, **15**, 661–75.

L, K.C. (1986). Asymptotic Optimality of CL and Generalized Cross-Validation in Ridge Regression with Application to Spline Smoothing. *Ann. Statist.*, **10**, 1101-12.

Pollard, D. (1986). Rates of uniform almost-sure convergence for empirical processes indexed by unbounded classes of functions, *Technical Report, Yale University Statistics Department.*

Pollard, D. (1995). Uniform Ratio Limit Theorems for Empirical Processes. *Scand. J. Stat.* , **22**, 271–78.

Rice, J. and Rosenblatt, M. (1983). Smoothing Splines: Regression, Derivatives and Deconvolution. *Ann. Statist.*, **11**, 141–56.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis.* Cambridge University press.

van der Geer, S. A. (2000). *Empirical Processes in M-Estimation.* Cambridge University Press, New York.

Wahba, G. (1990). *Spline models for observational data.* SIAM:Philadelphia.

Wood, S.N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Am. Statist. Assoc.*, **99**, 673–86.

Xiang, D. and Wahba, G. (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statistica Sinica*, **6**, 675–92.

Xu, G. and Huang, J.Z. (2012). Asymptotic optimality and efficient computation of the leave-subject-out cross-validation. *Ann. Statist.*, **40**, 3003-30.

Zhang, Y.C., Duchi, J., and Wainwright, M. (2015). Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.*, **16**, 3299–340.

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 $\mathbf{D}_1 = \mathrm{diag}\{\mathbf{B}_{11}, \ldots, \mathbf{B}_{mm}\}$ 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

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

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

The last inequality follows from condition C1. □

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

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

Proof. Using similar notations in equation (12), it is straightforward to show that

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

Using (12), we have that

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

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

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

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

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

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

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

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

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

where $\bar{b}_{ii}$'s are diagonal elements 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-C3 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 (13), we have that

$$\bar{U}(\lambda|\boldsymbol{X})-\bar{L}(\lambda|\boldsymbol{X})-\frac{1}{N}\boldsymbol{\varepsilon}^T\mathbf{W}\boldsymbol{\varepsilon} = \frac{2}{N}\boldsymbol{F}^T(\mathbf{I}-\bar{\mathbf{A}}_m)^T\mathbf{W}\boldsymbol{\varepsilon}-\frac{2}{N}\left\{\boldsymbol{\varepsilon}^T\bar{\mathbf{A}}_m\mathbf{W}\boldsymbol{\varepsilon}-\sigma^2\mathrm{tr}(\bar{\mathbf{A}}_m\mathbf{W})\right\}. \quad (A.5)$$

Notice that the random error ε and the covariate X are independent in model (1). We will show (16) using equation (A.1) in Lemma A.2, for which it suffices to show the following two equations

$$\mathrm{Var}_\varepsilon\left\{\frac{1}{N}\boldsymbol{F}^T(\mathbf{I}-\bar{\mathbf{A}}_m)^T\mathbf{W}\boldsymbol{\varepsilon}\right\} = o_{\mathbb{P}_X}\{\bar{R}^2(\lambda|\boldsymbol{X})\}, \quad (A.6)$$

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

We first show (A.6). Straightforward algebra yields that

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

where the second last equation follows from conditions C2-C3. Next, we show (A.7). Using condition C2, it is straightforward to show that

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

where $\bar{a}_{ii}$'s are diagonal elements of $\bar{\mathbf{A}}_m\mathbf{W}$ and $K_1 = \mathbb{E}_\varepsilon\varepsilon^4+\sigma^4$ is bounded. Hence, (A.7) is proved using, again, condition C2-C3. Using (A.6)-(A.7) and (A.1), the equation (16) 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

$$\mathrm{dGCV}_{DC}(\lambda|\boldsymbol{X}) - \bar{U}(\lambda|\boldsymbol{X}) = o_{\mathbb{P}_{\varepsilon,X}}\{\bar{R}(\lambda|\boldsymbol{X})\}. \tag{A.8}$$

Using the first order Taylor expansion of $(1-x)^{-2}$ around $x=0$, we have that $(1-x)^{-2} = 1+2x+6(1-x^*)^{-4}x^2$ 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

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

Using condition C4, we have that

$$\frac{\mathrm{tr}(\bar{\mathbf{A}}_m\mathbf{W})}{N} = o_{\mathbb{P}_X}\{\bar{R}^{1/2}(\lambda|\boldsymbol{X})\}, \tag{A.9}$$

which implies that $II = o_{\mathbb{P}_X}(\bar{R}(\lambda|\boldsymbol{X}))$ since $\frac{1}{N}\boldsymbol{Y}^T\{\mathbf{I}-\bar{\mathbf{A}}_m(\lambda)\}^T\mathbf{W}\{\mathbf{I}-\bar{\mathbf{A}}_m(\lambda)\}\boldsymbol{Y}$ is bounded. For part I, we can write

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

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

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

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 $N\bar{R}(\lambda|\boldsymbol{X})\xrightarrow{\mathbb{P}_X}\infty$ (condition C2), we have that $\frac{1}{N}\boldsymbol{\varepsilon}^T\mathbf{W}\boldsymbol{\varepsilon}-\sigma^2=o_{\mathbb{P}_{\varepsilon,X}}\{\bar{R}^{1/2}(\lambda|\boldsymbol{X})\}$. Using this and equation (A.9), we have that

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

The third part of I is $o_{\mathbb{P}_X}\{\bar{R}(\lambda|\boldsymbol{X})\}$ due to equation (A.9). Therefore, we have shown that

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

which completes the proof. □

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

$$\mathcal{F}=\{f\in L_2(\mathbb{P}):f\geq 0,\|f\|_{\sup}\leq V,J_1(f)\leq V^2H^2\},\tag{A.10}$$

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

$$\epsilon_n^3\gamma_n^2\geq\frac{c_0(1+H)V}{n},\tag{A.11}$$

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

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

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

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

□

Proof of Lemma 2. We first consider $Q_2(\lambda|\boldsymbol{X})$ in (20). 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.11) under Conditions C4'(b)-(c), we have that

$$P\left(\sup_{f\in\mathcal{V}_n}\frac{|\mathbb{P}_nf-\mathbb{P}f|}{\mathbb{P}_nf+\mathbb{P}f+\gamma_n} > C\epsilon\right) \leq \exp(-n\epsilon^2\gamma_n/2). \tag{A.12}$$

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

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

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 $Q(\lambda|\boldsymbol{X})$ in (18) and Q_j for $Q_j(\lambda|\boldsymbol{X})$, $j = 1, 2$, in (19) and (20) 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.12) 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

$$Q_2(\lambda|\boldsymbol{X}) = Q(\lambda|\boldsymbol{X}) + o_{\mathbb{P}_X} Q\{(\lambda|\boldsymbol{X})\}. \tag{A.13}$$

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

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

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.11) under Conditions C4'(b)-(c), we have that

$$P\left(\sup_{f\in\mathcal{V}_N} \frac{|\mathbb{P}_N f - \mathbb{P}f|}{\mathbb{P}_N f + \mathbb{P}f + \gamma_N} > C\epsilon\right) \leq \exp(-N\epsilon^2\gamma_N/2). \tag{A.14}$$

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

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

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

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

and that

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

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.14) gives that

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

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

$$Q_1(\lambda|\boldsymbol{X}) = Q(\lambda|\boldsymbol{X}) + o_{\mathbb{P}_X}\{Q(\lambda|\boldsymbol{X})\}. \tag{A.15}$$

Combining equations (A.13)–(A.15), we have that

$$\frac{\frac{1}{Nm}\sum_{k=1}^{m}\mathrm{tr}(\mathbf{A}_{kk}^2)}{\frac{\mathrm{tr}(\bar{\mathbf{A}}_m^T\bar{\mathbf{A}}_m)}{N}} = \frac{Q_1(\lambda|\boldsymbol{X})}{Q_2(\lambda|\boldsymbol{X})} = O_{\mathbb{P}_X}(1). \tag{A.16}$$

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.16), 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. □