Optimal Tuning for Divide-and-conquer Kernel Ridge Regression with Massive Data

Ganggang Xu [1] **Zuofeng Shang** [2] **Guang Cheng** [3]

1. Technical Proofs

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 S.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

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

which implies that

$$\begin{aligned}\lambda_{\max}(\bar{\mathbf{A}}_m\bar{\mathbf{A}}_m^T) &\le \frac{1}{m^2}\sum_{l=1}^m \lambda_{\max}\{\begin{pmatrix} \mathbf{K}_{1l} \\ \mathbf{K}_{2l} \\ \vdots \\ \mathbf{K}_{ml} \end{pmatrix}\mathbf{B}_{ll}(\mathbf{K}_{1l}^T,\ldots,\mathbf{K}_{ml}^T)\} \\ &= \frac{1}{m^2}\sum_{l=1}^m \lambda_{\max}(\mathbf{B}_{ll}\sum_{k=1}^m \mathbf{K}_{kl}^T\mathbf{K}_{kl}) \\ &= \frac{1}{m}\sum_{l=1}^m \lambda_{\max}\left\{(\mathbf{K}_{ll} + n_l\lambda\mathbf{I}_l)^{-2}\left(\frac{1}{m}\sum_{k=1}^m \mathbf{K}_{kl}^T\mathbf{K}_{kl}\right)\right\} \\ &= O_{\mathbb{P}_X}(1).\end{aligned}$$

The last inequality follows from condition C1. □

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

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

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

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

where $\boldsymbol{Y} = \boldsymbol{F} + \boldsymbol{\varepsilon}$. Using (12), we have that

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

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

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

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

*Equal contribution [1]Department of Mathematical Sciences, Binghamton University, the State University of New York, Binghamton, NY, USA [2]Department of Mathematical Sciences, Indiana UniversityPurdue University Indianapolis, IN, USA [3]Department of Statistics, Purdue University,West Lafayette, IN, USA . Correspondence to: Ganggang Xu <gang@math.binghamton.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

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

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

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

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

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

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-C3 and Lemma S.1. Using (S.3)-(S.4), the equation (S.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 (S.2) and (13), we have that

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

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

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

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

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

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

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

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

where $K_1=\mathbb{E}_{\varepsilon}\varepsilon^4+\sigma^4$ is bounded. Hence, (S.8) is proved using, again, condition C2-C3. Using (S.7)-(S.8) and (S.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 S.2, it suffices to show that

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

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 $\frac{\operatorname{tr}(\bar{\mathbf{A}}_m)}{N}=o_{\mathbb{P}_X}(1)$ and thus we can consider the following decomposition

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

Using condition C4, we have that

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

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

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

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 $\frac{\text{tr}(\bar{\mathbf{A}}_m\mathbf{W})}{N} = o_{\mathbb{P}_X}(1)$, and thus

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

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 (S.10), we have that

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

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

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

which completes the proof. □