Appendix for "Flexible Expectile Regression in Reproducing Kernel Hilbert Spaces"

In this appendix we provide technical proofs for the theorems and lemmas in "Flexible Expectile Regression in Reproducing Kernel Hilbert Spaces".

Some technical lemmas for Theorem 1

We first present some technical lemmas and their proofs. These lemmas are used to prove Theorem 1.

Lemma 3. *Let ϕ^*_ω be the convex conjugate of ϕ_ω,*

$$\phi^*_\omega(t) = \begin{cases} \frac{1}{4(1-\omega)} t^2 & \text{if } t \leq 0, \\ \frac{1}{4\omega} t^2 & \text{if } t > 0. \end{cases}$$

The solution to (10) *can be alternatively obtained by solving the optimization problem*

$$\min_{\{\alpha_i\}_{i=0}^n} g(\alpha_1, \alpha_2, \ldots, \alpha_n), \quad \text{subject to} \quad \sum_{i=1}^n \alpha_i = 0, \tag{29}$$

where g is defined by

$$g(\alpha_1, \alpha_2, \ldots, \alpha_n) = -\sum_{i=1}^n y_i \alpha_i + \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) + 2\lambda \sum_{i=1}^n \phi^*_\omega(\alpha_i). \tag{30}$$

Proof. Let $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)^\intercal$. Since both objective functions in (10) and (29) are convex, we only need to show that they share a common stationary point. Define

$$G_\omega(\boldsymbol{\alpha}) = \phi_\omega(\alpha_1) + \phi_\omega(\alpha_2) + \cdots + \phi_\omega(\alpha_n),$$

$$\nabla G_\omega(\boldsymbol{\alpha}) = (\phi'_\omega(\alpha_1), \phi'_\omega(\alpha_2), \ldots, \phi'_\omega(\alpha_n))^\intercal.$$

By setting the derivatives of (10) with respect to $\boldsymbol{\alpha}$ to be zero, we can find the stationary point of (10) satisfying

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\alpha}}\left[\begin{pmatrix} y_1-\alpha_0 \\ y_2-\alpha_0 \\ \vdots \\ y_n-\alpha_0 \end{pmatrix} - \mathbf{K}\boldsymbol{\alpha}\right]\cdot\left[\begin{array}{c} \phi'_\omega\big(y_1-\alpha_0-\sum_{j=1}^n K(x_1,x_j)\alpha_j\big) \\ \phi'_\omega\big(y_2-\alpha_0-\sum_{j=1}^n K(x_2,x_j)\alpha_j\big) \\ \vdots \\ \phi'_\omega\big(y_n-\alpha_0-\sum_{j=1}^n K(x_n,x_j)\alpha_j\big) \end{array}\right] + \lambda\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\alpha}}\boldsymbol{\alpha}^\intercal K\boldsymbol{\alpha} = \mathbf{0},$$

which can be reduced to

$$-\phi'_\omega\big(y_i-\alpha_0-\sum_{j=1}^n K(x_i,x_j)\alpha_j\big)+2\lambda\alpha_i=0, \quad \text{for } 1\le i\le n, \tag{31}$$

and setting the derivative of (10) with respect to α_0 to be zero, we have

$$\sum_{i=1}^n \phi'_\omega\big(y_i-\alpha_0-\sum_{j=1}^n K(x_i,x_j)\alpha_j\big)=0. \tag{32}$$

Combining (31) and (32), (32) can be simplified to

$$\sum_{i=1}^n \alpha_i = 0. \tag{33}$$

In comparison, the Lagrange function of (29) is

$$g(\alpha_1,\alpha_2,\ldots,\alpha_n)+\nu\sum_{i=1}^n\alpha_i. \tag{34}$$

The first order conditions of (34) are

$$-y_i+\nu+\sum_{j=1}^n K(x_i,x_j)\alpha_j+2\lambda{\phi^*_\omega}'(\alpha_i)=0, \quad \text{for } 1\le i\le n, \tag{35}$$

and

$$\sum_{i=1}^n \alpha_i = 0. \tag{36}$$

Noting that $2\lambda{\phi^*_\omega}'(\alpha_i)={\phi^*_\omega}'(2\lambda\alpha_i)$ and ${\phi^*_\omega}'$ is the inverse function of ϕ'_ω. Let $\nu=\alpha_0$, then (31) and (35) are equivalent. Therefore, (10) and (29) have a common stationary point and therefore a common minimizer. □

Lemma 4.

$$\sum_{j=1}^{n} \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \leq \sqrt{K(\mathbf{x}_i, \mathbf{x}_i)} \cdot \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)}.$$

Proof. Let $\mathbf{C} = \mathbf{K}^{1/2}$, then by Cauchy-Schwarz inequality

$$\sum_{j=1}^{n} \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) = (\alpha_1, \alpha_2, \ldots, \alpha_n)\mathbf{C}(\mathbf{C}_{i,1}, \mathbf{C}_{i,2}, \ldots, \mathbf{C}_{i,n})^T$$

$$\leq \|(\alpha_1, \alpha_2, \ldots, \alpha_n)\mathbf{C}\| \cdot \|(\mathbf{C}_{i,1}, \mathbf{C}_{i,2}, \ldots, \mathbf{C}_{i,n})\| = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)} \cdot \sqrt{K(\mathbf{x}_i, \mathbf{x}_i)}.$$

□

Lemma 5. *For the g function defined in* (30)*, we have*

$$\begin{aligned} &\frac{1}{2} \sum_{i,j=1}^{n} (\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j) K(\mathbf{x}_i, \mathbf{x}_j) + \frac{\lambda}{2\max(1-\omega, \omega)} \sum_{i=1}^{n} (\alpha_i - \hat{\alpha}_i)^2 \\ \leq & g(\alpha_1, \alpha_2, \ldots, \alpha_n) - g(\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_n) \\ \leq & \frac{1}{2} \sum_{i,j=1}^{n} (\alpha_i - \hat{\alpha}_i)(\alpha_j - \hat{\alpha}_j) K(\mathbf{x}_i, \mathbf{x}_j) + \frac{\lambda}{2\min(1-\omega, \omega)} \sum_{i=1}^{n} (\alpha_i - \hat{\alpha}_i)^2. \end{aligned}$$

Proof. It is clear that the second derivative of g is bounded above by $\mathbf{K} + \frac{\lambda}{\min(1-\omega,\omega)}\mathbf{I}$ and bounded below by $\mathbf{K} + \frac{\lambda}{\max(1-\omega,\omega)}\mathbf{I}$, where $\mathbf{K} \in \mathbb{R}^{n,n}$. Let $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)^\intercal$

$$g(\boldsymbol{\alpha}) - g(\widehat{\boldsymbol{\alpha}}) \leq g'(\widehat{\boldsymbol{\alpha}})^\intercal(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}) + \frac{1}{2}(\mathbf{K} + \frac{\lambda}{\min(1-\omega, \omega)}\mathbf{I})(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}})^\intercal(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}), \tag{37}$$

$$g(\boldsymbol{\alpha}) - g(\widehat{\boldsymbol{\alpha}}) \geq g'(\widehat{\boldsymbol{\alpha}})^\intercal(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}) + \frac{1}{2}(\mathbf{K} + \frac{\lambda}{\max(1-\omega, \omega)}\mathbf{I})(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}})^\intercal(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}). \tag{38}$$

Hence when $\boldsymbol{\alpha}$ and $\widehat{\boldsymbol{\alpha}}$ are fixed and $g'(\widehat{\boldsymbol{\alpha}}) = 0$, the maximum of $g(\boldsymbol{\alpha}) - g(\widehat{\boldsymbol{\alpha}})$ is obtained when the second order derivative of g achieves its maximum and the minimum is obtained when the second order derivative achieves its minimum. □

The next lemma establishes the basis for the so-called leave-one-out analysis (Jaakkola and Haussler, 1999; Joachims, 2000; Forster and Warmuth, 2002; Zhang, 2003). The basic idea is that the expected observed risk is equivalent to the expected leave-one-out error. Let $D_{n+1} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n+1}$ be a random sample of size $n+1$, and let $D_{n+1}^{[i]}$ be the subset of D_{n+1} with the i-th observation removed, i.e.

$$D_{n+1}^{[i]} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{i-1}, y_{i-1}), (\mathbf{x}_{i+1}, y_{i+1}), \ldots, (\mathbf{x}_{n+1}, y_{n+1})\}.$$

Let $(\hat{f}^{[i]}, \hat{\alpha}_0^{[i]})$ be the estimator trained on $D_{n+1}^{[i]}$. The leave-one-out error is defined as the averaged prediction error on each observation $(\mathbf{x}_i, y_i)$ using the estimator $(\hat{f}^{[i]}, \hat{\alpha}_0^{[i]})$ computed from $D_{n+1}^{[i]}$, where $(\mathbf{x}_i, y_i)$ is excluded:

$$\text{Leave-one-out error:} \quad \frac{1}{n+1} \sum_{i=1}^{n+1} \phi_\omega(y_i - \hat{\alpha}_0^{[i]} - \hat{f}^{[i]}(\mathbf{x}_i)).$$

Lemma 6. *Let $(\hat{f}_{(n)}, \hat{\alpha}_{0\,(n)})$ be the KERE estimator trained from D_n. The expected observed risk $E_{D_n} E_{(\mathbf{x},y)} \phi_\omega(y - \hat{\alpha}_{0\,(n)} - \hat{f}_{(n)}(\mathbf{x}))$ is equivalent to the expected leave-one-out error on D_{n+1}:*

$$E_{D_n}\big\{E_{(\mathbf{x},y)} \phi_\omega(y - \hat{\alpha}_{0\,(n)} - \hat{f}_{(n)}(\mathbf{x}))\big\} = E_{D_{n+1}}\Big(\frac{1}{n+1} \sum_{i=1}^{n+1} \phi_\omega(y_i - \hat{\alpha}_0^{[i]} - \hat{f}^{[i]}(\mathbf{x}_i))\Big), \tag{39}$$

where $\hat{\alpha}_0^{[i]}$ and $\hat{f}^{[i]}$ are KERE trained from $D_{n+1}^{[i]}$.

Proof.

$$\begin{aligned}
E_{D_{n+1}}\Big(\frac{1}{n+1} \sum_{i=1}^{n+1} \phi_\omega(y_i - \hat{\alpha}_0^{[i]} - \hat{f}^{[i]}(\mathbf{x}_i))\Big) &= \frac{1}{n+1} \sum_{i=1}^{n+1} E_{D_{n+1}} \phi_\omega(y_i - \hat{\alpha}_0^{[i]} - \hat{f}^{[i]}(\mathbf{x}_i)) \\
&= \frac{1}{n+1} \sum_{i=1}^{n+1} E_{D_{n+1}^{[i]}}\big\{E_{(\mathbf{x}_i, y_i)} \phi_\omega(y_i - \hat{\alpha}_0^{[i]} - \hat{f}^{[i]}(\mathbf{x}_i))\big\} \\
&= \frac{1}{n+1} \sum_{i=1}^{n+1} E_{D_n}\big\{E_{(\mathbf{x},y)} \phi_\omega(y - \hat{\alpha}_0 - \hat{f}(\mathbf{x}))\big\} \\
&= E_{D_n} E_{(\mathbf{x},y)} \phi_\omega(y - \hat{\alpha}_0 - \hat{f}(\mathbf{x})).
\end{aligned}$$

□

In the following Lemma, we give an upper bound of $|\hat{\alpha}_i|$ for $1 \le i \le n$.

Lemma 7. *Assume* $M = \sup_{\mathbf{x}} K(\mathbf{x},\mathbf{x})^{1/2}$*. Denote as* $(\hat{f}_{(n)}, \hat{\alpha}_{0\,(n)})$ *the KERE estimator in* (7) *trained on* n *samples* $D_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$*. The estimates* $\hat{\alpha}_{i\,(n)}$ *for* $1 \le i \le n$ *are defined by* $\hat{f}_{(n)}(\cdot) = \sum_{i=1}^n \hat{\alpha}_{i\,(n)} K(\mathbf{x}_i, \cdot)$*. Denote* $\|Y_n\|_2 = \sqrt{\sum_{i=1}^n y_i^2}$*,* $\frac{\|Y_n\|_1}{n} = \frac{1}{n}\sum_{i=1}^n |y_i|$*,* $q_1 = \frac{\max(1-\omega,\omega)}{\min(1-\omega,\omega)}$*,* $q_2 = \max(1-\omega,\omega)$*. We claim that*

$$|\hat{\alpha}_{i\,(n)}| \le \frac{q_2}{\lambda}\Big(q_1 \frac{\|Y_n\|_1}{n} + M(q_1+1)\sqrt{\frac{q_2}{\lambda}}\|Y_n\|_2 + |y_i|\Big), \quad \textit{for } 1 \le i \le n. \tag{40}$$

Proof. The proof is as follows. The function g is defined as in (30), then

$$g(\hat{\alpha}_{1\,(n)}, \hat{\alpha}_{2\,(n)}, \ldots, \hat{\alpha}_{n\,(n)}) \le g(0, 0, \ldots, 0) = 0,$$

we have

$$\begin{aligned}
\frac{1}{2}\sum_{i,j=1}^n \hat{\alpha}_{i\,(n)}\hat{\alpha}_{j\,(n)} K(\mathbf{x}_i, \mathbf{x}_j) &\le \sum_{i=1}^n y_i \hat{\alpha}_{i\,(n)} - 2\lambda \sum_{i=1}^n \phi_\omega^*(\hat{\alpha}_{i\,(n)}) \\
&\le -\frac{\lambda}{2q_2}\sum_{i=1}^n \Big(\hat{\alpha}_{i\,(n)} - \frac{q_2}{\lambda} y_i\Big)^2 + \frac{q_2}{2\lambda}\sum_{i=1}^n y_i^2 \\
&\le \frac{q_2}{2\lambda}\sum_{i=1}^n y_i^2.
\end{aligned}$$

Applying Lemma 4, we have

$$\hat{f}_{(n)}(\mathbf{x}_i) = \sum_{j=1}^n \hat{\alpha}_{j\,(n)} K(\mathbf{x}_i, \mathbf{x}_j) \le M\sqrt{\frac{q_2 \sum_{i=1}^n y_i^2}{\lambda}} = M\sqrt{\frac{q_2}{\lambda}}\|Y_n\|_2. \tag{41}$$

By the definition in (10), $\hat{\alpha}_{0\,(n)}$ is given by $\operatorname{argmin}_{\alpha_0} \sum_{i=1}^n \phi_\omega\big(y_i - \alpha_0 - \hat{f}_{(n)}(\mathbf{x}_i)\big)$. By the first order condition

$$\sum_{i=1}^n 2\big|\omega - I(y_i - \hat{\alpha}_{0\,(n)} - \hat{f}_{(n)}(\mathbf{x}_i))\big|(y_i - \hat{\alpha}_{0\,(n)} - \hat{f}_{(n)}(\mathbf{x}_i)) = 0.$$

Let $c_i = \big|\omega - I(y_i - \hat{\alpha}_{0\,(n)} - \hat{f}_{(n)}(\mathbf{x}_i))\big|$, we have $\min(1-\omega, \omega) \leq c_i \leq \max(1-\omega, \omega)$, hence

$$\begin{aligned}\Big|\Big(\sum_{i=1}^{n} c_i\Big)\hat{\alpha}_{0\,(n)}\Big| &= \Big|\sum_{i=1}^{n} c_i(y_i - \hat{f}_{(n)}(\mathbf{x}_i))\Big| \leq \sum_{i=1}^{n} c_i(\big|y_i\big| + \big|\hat{f}_{(n)}(\mathbf{x}_i)\big|) \\ &\leq q_2\Big(\sum_{i=1}^{n}|y_i| + nM\sqrt{\frac{q_2}{\lambda}}\|Y_n\|_2\Big),\end{aligned}$$

and we have

$$|\hat{\alpha}_{0\,(n)}| \leq q_1\Big(\frac{\|Y_n\|_1}{n} + M\sqrt{\frac{q_2}{\lambda}}\|Y_n\|_2\Big). \tag{42}$$

Combining (31) and (42), we concluded (40). □

Proof of Theorem 1

Proof. Consider $n+1$ training samples $D_{n+1} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n+1}, y_{n+1})\}$. Denote as $(\hat{f}^{[i]}, \hat{\alpha}_0^{[i]})$ the KERE estimator trained from $D_{n+1}^{[i]}$, which is a subset of D_{n+1} with i-th observation removed, i.e.,

$$D_{n+1}^{[i]} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{i-1}, y_{i-1}), (\mathbf{x}_{i+1}, y_{i+1}), \ldots, (\mathbf{x}_{n+1}, y_{n+1})\}.$$

Denote as $(\hat{f}_{(n+1)}, \hat{\alpha}_{0\,(n+1)})$ the KERE estimator trained from $n+1$ samples D_{n+1}. The estimates $\hat{\alpha}_i$ for $1 \leq i \leq n+1$ are defined by $\hat{f}_{(n+1)}(\cdot) = \sum_{i=1}^{n+1}\hat{\alpha}_i K(\mathbf{x}_i, \cdot)$.

In what follows, we denote $\|Y_{n+1}\|_2 = \sqrt{\sum_{i=1}^{n+1} y_i^2}$, $\frac{\|Y_{n+1}\|_1}{n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1}|y_i|$, $q_1 = \frac{\max(1-\omega,\omega)}{\min(1-\omega,\omega)}$, $q_2 = \max(1-\omega, \omega)$, $q_3 = \min(1-\omega, \omega)$.

Part I We first show that the leave-one-out estimate is sufficiently close to the estimate fitted from using all the training data. Without loss of generality, just consider the case that the $(n+1)$th data point is removed. The same results apply to the other leave-one out cases. We show that $|\hat{f}^{[n+1]}(\mathbf{x}_i) + \hat{\alpha}_0^{[n+1]} - \hat{f}_{(n+1)}(\mathbf{x}_i) - \hat{\alpha}_{0\,(n+1)}| \leq C_2^{[n+1]}$, where the expression of $C_2^{[n+1]}$ is to be derived in the following.

We first study the upper bound for $|\hat{f}^{[n+1]}(\mathbf{x}_i) - \hat{f}_{(n+1)}(\mathbf{x}_i)|$. By the definitions of g in

(30) and $(\hat{\alpha}_1^{[n+1]}, \hat{\alpha}_2^{[n+1]}, \ldots, \hat{\alpha}_n^{[n+1]})$, we have

$$\begin{aligned} & g\big(\hat{\alpha}_1^{[n+1]}, \hat{\alpha}_2^{[n+1]}, \ldots, \hat{\alpha}_n^{[n+1]}, 0\big) \\ = \;& g\big(\hat{\alpha}_1^{[n+1]}, \hat{\alpha}_2^{[n+1]}, \ldots, \hat{\alpha}_n^{[n+1]}\big) \\ \leq \;& g\big(\hat{\alpha}_1 + \frac{1}{n}\hat{\alpha}_{n+1}, \hat{\alpha}_2 + \frac{1}{n}\hat{\alpha}_{n+1}, \ldots, \hat{\alpha}_n + \frac{1}{n}\hat{\alpha}_{n+1}\big) \\ = \;& g\big(\hat{\alpha}_1 + \frac{1}{n}\hat{\alpha}_{n+1}, \hat{\alpha}_2 + \frac{1}{n}\hat{\alpha}_{n+1}, \ldots, \hat{\alpha}_n + \frac{1}{n}\hat{\alpha}_{n+1}, 0\big). \end{aligned}$$

That is,

$$\begin{aligned} & g\big(\hat{\alpha}_1^{[n+1]}, \hat{\alpha}_2^{[n+1]}, \ldots, \hat{\alpha}_n^{[n+1]}, 0\big) - g\big(\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{n+1}\big) \\ & \leq g\big(\hat{\alpha}_1 + \frac{1}{n}\hat{\alpha}_{n+1}, \hat{\alpha}_2 + \frac{1}{n}\hat{\alpha}_{n+1}, \ldots, \hat{\alpha}_n + \frac{1}{n}\hat{\alpha}_{n+1}, 0\big) - g\big(\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{n+1}\big). \end{aligned}$$

Denote for simplicity that $\hat{\alpha}_{n+1}^{[n+1]} = 0$. Applying Lemma 5 to both LHS and RHS of the above inequality, we have

$$\begin{aligned} & \sum_{i,j=1}^{n+1} (\hat{\alpha}_i^{[n+1]} - \hat{\alpha}_i)(\hat{\alpha}_j^{[n+1]} - \hat{\alpha}_j) K(\mathbf{x}_i, \mathbf{x}_j) + \frac{\lambda}{2q_2} \sum_{i=1}^{n+1} (\hat{\alpha}_i^{[n+1]} - \hat{\alpha}_i)^2 \\ \leq & \hat{\alpha}_{n+1}^2 \Big[\Big(\frac{1}{n}, \ldots, \frac{1}{n}, -1\Big) \mathbf{K} \Big(\frac{1}{n}, \ldots, \frac{1}{n}, -1\Big)^T + \frac{\lambda(n+1)}{2nq_3} \Big], \end{aligned}$$

where $\mathbf{K} \in \mathbb{R}^{n+1,n+1}$ is defined by $\mathbf{K}_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j)$. Since $|K(\mathbf{x}_i, \mathbf{x}_j)| \leq M^2$ for any $1 \leq i, j \leq n+1$, we have

$$\begin{aligned} & \Big(\tfrac{1}{n}, \ldots, \tfrac{1}{n}, -1\Big) \mathbf{K} \Big(\tfrac{1}{n}, \ldots, \tfrac{1}{n}, -1\Big)^T \\ = \;& \tfrac{1}{n^2} \textstyle\sum_{i,j=1}^n \mathbf{K}_{i,j} - \tfrac{1}{n} \sum_{i=1}^n \mathbf{K}_{i,n+1} - \tfrac{1}{n} \sum_{j=1}^n \mathbf{K}_{n+1,j} + \mathbf{K}_{n+1,n+1} \\ \leq \;& M^2 + M^2 + M^2 + M^2 = 4M^2. \end{aligned}$$

Combining it with the bound for $|\hat{\alpha}_{n+1}|$ by Lemma 7 (note that here $\hat{\alpha}_{n+1}$ is trained on $n+1$ samples), we have

$$\sum_{i,j=1}^{n+1} (\hat{\alpha}_i^{[n+1]} - \hat{\alpha}_i)(\hat{\alpha}_j^{[n+1]} - \hat{\alpha}_j) K(\mathbf{x}_i, \mathbf{x}_j) \leq C_1^{[n+1]}, \tag{43}$$

where

$$C_1^{[n+1]} = \left(4M^2 + \frac{\lambda(n+1)}{2nq_3}\right)\left(\frac{q_2}{\lambda}C_0^{[n+1]}\right)^2, \tag{44}$$

and

$$C_0^{[n+1]} = q_1\frac{\|Y_{n+1}\|_1}{n+1} + M(q_1+1)\sqrt{\frac{q_2}{\lambda}}\|Y_{n+1}\|_2 + |y_{n+1}|. \tag{45}$$

Combining (43) with Lemma 4, we have that for $1 \le i \le n+1$,

$$|\hat{f}^{[n+1]}(\mathbf{x}_i) - \hat{f}_{(n+1)}(\mathbf{x}_i)| = \Big|\sum_{j=1}^{n+1}(\hat{\alpha}_i^{[n+1]} - \hat{\alpha}_i)K(\mathbf{x}_i,\mathbf{x}_j)\Big| \le \sqrt{C_1^{[n+1]}}M. \tag{46}$$

Next, we bound $|\hat{\alpha}_0^{[n+1]} - \hat{\alpha}_{0\,(n+1)}|$. Since $\hat{\alpha}_{0\,(n+1)}$ and $\hat{\alpha}_0^{[n+1]}$ are the minimizers of

$$\sum_{i=1}^{n+1}\phi_\omega\left(y_i - \alpha_0 - \hat{f}_{(n+1)}(\mathbf{x}_i)\right) \text{ and } \sum_{i=1}^{n}\phi_\omega\left(y_i - \alpha_0 - \hat{f}^{[n+1]}(\mathbf{x}_i)\right),$$

we have

$$\frac{\mathrm{d}}{\mathrm{d}\alpha_0}\sum_{i=1}^{n+1}\phi_\omega\left(y_i - \alpha_0 - \hat{f}_{(n+1)}(\mathbf{x}_i)\right)\Big|_{\alpha_0=\hat{\alpha}_{0\,(n+1)}} = 0, \tag{47}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}\alpha_0}\sum_{i=1}^{n}\phi_\omega\left(y_i - \alpha_0 - \hat{f}^{[n+1]}(\mathbf{x}_i)\right)\Big|_{\alpha_0=\hat{\alpha}_0^{[n+1]}} = 0. \tag{48}$$

By the Lipschitz continuity of ϕ'_ω we have

$$\begin{aligned}&\left|\textstyle\sum_{i=1}^{n+1}\phi'_\omega\left(y_i - \hat{\alpha}_{0\,(n+1)} - \hat{f}^{[n+1]}(\mathbf{x}_i)\right) - \sum_{i=1}^{n+1}\phi'_\omega\left(y_i - \hat{\alpha}_{0\,(n+1)} - \hat{f}_{(n+1)}(\mathbf{x}_i)\right)\right| \\ \le\ & 2(n+1)q_2|\hat{f}^{[n+1]}(\mathbf{x}_i) - \hat{f}_{(n+1)}(\mathbf{x}_i)|,\end{aligned}$$

and by applying (46) and (47) we have the upper bound

$$\left|\sum_{i=1}^{n+1}\phi'_\omega\left(y_i - \hat{\alpha}_{0\,(n+1)} - \hat{f}^{[n+1]}(\mathbf{x}_i)\right)\right| \le 2(n+1)q_2\sqrt{C_1^{[n+1]}}M.$$

Similarly, by (41), (42), and (48) we have

$$\begin{aligned}
& \left| \textstyle\sum_{i=1}^{n} \phi'_\omega \left(y_i - \hat{\alpha}_{0\,(n+1)} - \hat{f}^{[n+1]}(\mathbf{x}_i) \right) \right| \\
= & \left| \textstyle\sum_{i=1}^{n+1} \phi'_\omega \left(y_i - \hat{\alpha}_{0\,(n+1)} - \hat{f}^{[n+1]}(\mathbf{x}_i) \right) - \textstyle\sum_{i=1}^{n+1} \phi'_\omega \left(y_i - \hat{\alpha}_{0\,(n+1)} - \hat{f}_{(n+1)}(\mathbf{x}_i) \right) \right. \\
& \left. - \phi'_\omega \left(y_{n+1} - \hat{\alpha}_{0\,(n+1)} - \hat{f}^{[n+1]}(\mathbf{x}_{n+1}) \right) \right| \\
\leq & \left| \textstyle\sum_{i=1}^{n+1} \phi'_\omega \left(y_i - \hat{\alpha}_{0\,(n+1)} - \hat{f}^{[n+1]}(\mathbf{x}_i) \right) - \textstyle\sum_{i=1}^{n+1} \phi'_\omega \left(y_i - \hat{\alpha}_{0\,(n+1)} - \hat{f}_{(n+1)}(\mathbf{x}_i) \right) \right| \qquad (49) \\
& + \left| \phi'_\omega \left(y_{n+1} - \hat{\alpha}_{0\,(n+1)} - \hat{f}^{[n+1]}(\mathbf{x}_{n+1}) \right) \right| \\
\leq & \; 2(n+1) q_2 \sqrt{C_1^{[n+1]}} M + 2 q_2 \left(|y_{n+1}| + |\hat{\alpha}_{0\,(n+1)}| + |\hat{f}_{(n)}| \right) \\
\leq & \; 2(n+1) q_2 \sqrt{C_1^{[n+1]}} M + 2 q_2 \left(|y_{n+1}| + q_1 \tfrac{\|Y_{n+1}\|_1}{n+1} + M q_1 \sqrt{\tfrac{q_2}{\lambda}} \|Y_{n+1}\|_2 + \sqrt{\tfrac{q_2}{\lambda}} \|Y_n\|_2 \right) \\
\leq & \; 2(n+1) q_2 \sqrt{C_1^{[n+1]}} M + 2 q_2 C_0^{[n+1]},
\end{aligned}$$

where the second last inequality follows from (41) and (42). Note that in this case the corresponding sample is $n+1$.

Using (48) we have

$$\begin{aligned}
& 2nq_3 \left| \hat{\alpha}_0^{[n+1]} - \hat{\alpha}_{0\,(n+1)} \right| \\
\leq & \left| \sum_{i=1}^{n} \phi'_\omega \left(y_i - \hat{\alpha}_0^{[n+1]} - \hat{f}^{[n+1]}(\mathbf{x}_i) \right) - \sum_{i=1}^{n} \phi'_\omega \left(y_i - \hat{\alpha}_{0\,(n+1)} - \hat{f}^{[n+1]}(\mathbf{x}_i) \right) \right| \\
= & \left| \sum_{i=1}^{n} \phi'_\omega \left(y_i - \hat{\alpha}_{0\,(n+1)} - \hat{f}^{[n+1]}(\mathbf{x}_i) \right) \right|.
\end{aligned}$$

By (49), we have

$$|\hat{\alpha}_0^{[n+1]} - \hat{\alpha}_{0\,(n+1)}| \leq q_1 \Big((1 + \frac{1}{n}) \sqrt{C_1^{[n+1]}} M + \frac{1}{n} C_0^{[n+1]} \Big). \qquad (50)$$

Finally, combining (46) and (50) we have

$$|\hat{f}^{[n+1]}(\mathbf{x}_i) + \hat{\alpha}_0^{[n+1]} - \hat{f}_{(n+1)}(\mathbf{x}_i) - \hat{\alpha}_{0\,(n+1)}| \leq C_2^{[n+1]}, \qquad (51)$$

where

$$C_2^{[n+1]} = q_1\Big((1+\frac{1}{n})\sqrt{C_1^{[n+1]}}M + \frac{1}{n}C_0^{[n+1]}\Big) + \sqrt{C_1^{[n+1]}}M. \tag{52}$$

Part II We now use (51) to derive a bound for $\phi_\omega(y_{n+1} - \hat{\alpha}_0^{[n+1]} - \hat{f}^{[n+1]}(\mathbf{x}_{n+1}))$. Let $t = \hat{f}_{(n+1)}(\mathbf{x}_i) + \hat{\alpha}_{0\,(n+1)} - \hat{f}^{[n+1]}(\mathbf{x}_i) - \hat{\alpha}_0^{[n+1]}$ and $t' = y_i - \hat{\alpha}_{0\,(n+1)} - \hat{f}_{(n+1)}(\mathbf{x}_i)$. We claim that,

$$\phi_\omega(t+t') - \phi_\omega(t') \le q_2(|2tt'| + |t^2|). \tag{53}$$

when $(t+t')$ and t' are both positive or both negative, (53) follows from $(t+t')^2 - t'^2 = 2tt' + t^2$. When $t+t'$ and t' have different signs, it must be that $|t'| < |t|$, and we have $|t| = |t+t'| + |t'|$ and hence $|t+t'| < |t|$. Then (53) is proved by $\phi_\omega(t+t') - \phi_\omega(t') = \max(\phi_\omega(t+t'), \phi_\omega(t')) \le q_2 \max((t+t')^2, t'^2) \le \max(1-\omega,\omega)t^2 < \max(1-\omega,\omega)(|2tt'| + |t^2|)$.

Hence by (51), (53) and the upper bound of $|y_{n+1} - \hat{f}_{(n+1)}(\mathbf{x}_{n+1}) - \hat{\alpha}_{0\,(n+1)}|$, we have

$$\phi_\omega(y_{n+1} - \hat{\alpha}_0^{[n+1]} - \hat{f}^{[n+1]}(\mathbf{x}_{n+1})) \le \phi_\omega(y_{n+1} - \hat{\alpha}_{0\,(n+1)} - \hat{f}_{(n+1)}(\mathbf{x}_{n+1})) + C_3^{[n+1]}, \tag{54}$$

where

$$C_3^{[n+1]} = q_2(2C_0^{[n+1]}C_2^{[n+1]} + (C_2^{[n+1]})^2). \tag{55}$$

Note that (54) and (55) hold for other $i, 1 \le i \le n+1$.

$$\phi_\omega(y_i - \hat{\alpha}_0^{[i]} - \hat{f}^{[i]}(\mathbf{x}_i)) \le \phi_\omega(y_i - \hat{\alpha}_{0\,(n+1)} - \hat{f}_{(n+1)}(\mathbf{x}_i)) + C_3^{[i]}. \tag{56}$$

Hence by (44), (45), (52) and (54) we have

$$E_{D_{n+1}}\Big(\phi_\omega(y_i - \hat{\alpha}_0^{[i]} - \hat{f}^{[i]}(\mathbf{x}_i))\Big) \le E_{D_{n+1}}\Big(\phi_\omega(y_i - \hat{\alpha}_{0\,(i)} - \hat{f}_{(n+1)}(\mathbf{x}_i))\Big) + E_{D_{n+1}}C_3^{[i]}. \tag{57}$$

and

$$\begin{aligned} & \frac{1}{n+1}E_{D_{n+1}}\Big(\sum_{i=1}^{n+1}\phi_\omega(y_i - \hat{\alpha}_0^{[i]} - \hat{f}^{[i]}(\mathbf{x}_i))\Big) \\ \le\ & \frac{1}{n+1}E_{D_{n+1}}\Big(\sum_{i=1}^{n+1}\phi_\omega(y_i - \hat{\alpha}_{0\,(n+1)} - \hat{f}_{(n+1)}(\mathbf{x}_i))\Big) + \frac{1}{n+1}E_{D_{n+1}}\sum_{i=1}^{n+1}C_3^{[i]}. \end{aligned} \tag{58}$$

On the other hand, let $(f_\varepsilon^*, \alpha_{0\,\varepsilon}^*)$ in the RKHS and satisfy $\mathcal{R}(f_\varepsilon^*, \alpha_{0\,\varepsilon}^*) \le \inf_{f\in\mathbb{H}_K, \alpha_0\in\mathbb{R}} \mathcal{R}(f, \alpha_0) + \varepsilon$. From the definition of $\hat{\alpha}_{0\,(n+1)}$ and $\hat{f}_{(n+1)}$ we have

$$
\begin{aligned}
&\frac{1}{n+1}\Big(\sum_{i=1}^{n+1}\phi_\omega(y_i - \hat{\alpha}_{0\,(n+1)} - \hat{f}_{(n+1)}(\mathbf{x}_i))\Big) + \frac{\lambda}{n+1}\|\hat{f}_{(n+1)}\|_{\mathbb{H}_K}^2 \\
\le\ &\frac{1}{n+1}\Big(\sum_{i=1}^{n+1}\phi_\omega(y_i - \alpha_{0\,\varepsilon}^* - f_\varepsilon^*(\mathbf{x}_i))\Big) + \frac{\lambda}{n+1}\|f_\varepsilon^*\|_{\mathbb{H}_K}^2.
\end{aligned} \tag{59}
$$

By Lemma 6, (58) and (59), we get

$$
\begin{aligned}
&E_{D_n}\big\{E_{(\mathbf{x},y)}\phi_\omega(y - \hat{\alpha}_{0\,(n)} - \hat{f}_{(n)}(\mathbf{x}))\big\} \\
=\ &\frac{1}{n+1}E_{D_{n+1}}\Big(\sum_{i=1}^{n+1}\phi_\omega(y_i - \hat{\alpha}_0^{[i]} - \hat{f}^{[i]}(\mathbf{x}_i))\Big) \\
\le\ &E_{D_n}\big\{E_{(\mathbf{x},y)}\phi_\omega(y - \alpha_{0\,\varepsilon}^* - f_\varepsilon^*(\mathbf{x}_i))\big\} + \frac{\lambda}{n+1}\|f_\varepsilon^*\|_{\mathbb{H}_K}^2 + \frac{1}{n+1}E_{D_{n+1}}\sum_{i=1}^{n+1}C_3^{[i]} \\
\le\ &\inf_{f\in\mathbb{H}_K, \alpha_0\in\mathbb{R}}\mathcal{R}(f, \alpha_0) + \varepsilon + \frac{\lambda}{n+1}\|f_\varepsilon^*\|_{\mathbb{H}_K}^2 + \frac{1}{n+1}E_{D_{n+1}}\sum_{i=1}^{n+1}C_3^{[i]}.
\end{aligned} \tag{60}
$$

Because $\lambda/n \to 0$, there exists N_ε such that when $n > N_\varepsilon$, $\frac{\lambda}{n+1}\|f_\varepsilon^*\|_{\mathbb{H}_K}^2 \le \varepsilon$. In what follows, we show that there exists N_ε' such that when $n > N_\varepsilon'$, $\frac{1}{n+1}E_{D_{n+1}}\sum_{i=1}^{n+1}C_3^{[i]} \le \varepsilon$. Thus, when $n > \max(N_\varepsilon, N_\varepsilon')$ we have

$$
E_{D_n}\big\{E_{(\mathbf{x},y)}\phi_\omega(y - \hat{\alpha}_{0\,(n)} - \hat{f}_{(n)}(\mathbf{x}))\big\} \le \inf_{f\in\mathbb{H}_K, \alpha_0\in\mathbb{R}}\mathcal{R}(f, \alpha_0) + 3\varepsilon.
$$

Since it holds for any $\varepsilon > 0$, Theorem 1 will be proved.

Now we only need to show that $\frac{1}{n+1}E_{D_{n+1}}\sum_{i=1}^{n+1}C_3^{[i]} \to 0$ as $n \to \infty$. In fact we can show $\frac{1}{n+1}E_{D_{n+1}}\sum_{i=1}^{n+1}C_3^{[i]} \le \frac{C}{\sqrt{\lambda}}D\left(\frac{1+n}{\lambda} + 1\right) \to 0$ as $n \to \infty$. In the following analysis, C represents any constant that does not depend on n, but the value of C varies in different expressions. Let $V_i = q_1\frac{\|Y_{n+1}\|_1}{n+1} + M(q_1+1)\sqrt{\frac{q_2}{\lambda}}\|Y_{n+1}\|_2 + |y_i|$, then as $n \to \infty$, $4M^2 < \frac{\lambda(n+1)}{2nq_3}$, and we have the upper bound

$$
C_1^{[i]} < (C\lambda)\Big(\frac{C}{\lambda}V_i\Big)^2 = C\frac{V_i^2}{\lambda},
$$

and since $n > \sqrt{\lambda}$ asymptotically, we have

$$C_2^{[i]} < C\Big(C\sqrt{C_1^{[i]}} + \frac{V_i}{n}\Big) + C\sqrt{C_1^{[i]}} < C\frac{V_i}{\sqrt{\lambda}} + C\frac{V_i}{n} < C\frac{V_i}{\sqrt{\lambda}}.$$

Then

$$C_3^{[i]} < CV_iC_2^{[i]} + CC_2^{[i]\,2} < CV_i\frac{V_i}{\sqrt{\lambda}} + C\frac{V_i^2}{\lambda} < C\frac{V_i^2}{\sqrt{\lambda}}. \tag{61}$$

We can bound V_i as follows:

$$\begin{aligned} V_i &= q_1\frac{\|Y_{n+1}\|_1}{n+1} + M(q_1+1)\sqrt{\frac{q_2}{\lambda}}\|Y_{n+1}\|_2 + |y_i| \\ &< q_1\frac{\|Y_{n+1}\|_2}{\sqrt{n+1}} + M(q_1+1)\sqrt{\frac{q_2}{\lambda}}\|Y_{n+1}\|_2 + |y_i| \\ &< C\sqrt{\frac{\|Y_{n+1}\|_2^2}{\lambda}} + C|y_i|. \end{aligned}$$

Then we have

$$E_{D_{n+1}}V_i^2 \;<\; 2C^2E_{D_{n+1}}\Big[\frac{\|Y_{n+1}\|_2^2}{\lambda} + y_i^2\Big]. \tag{62}$$

Combining it with (61) and using the assumption $Ey_i^2 < D$, we have

$$\begin{aligned} \frac{1}{n+1}E_{D_{n+1}}\sum_{i=1}^{n+1}C_3^{[i]} &\le \frac{C}{\sqrt{\lambda}}\frac{1}{1+n}\left(\frac{1+n}{\lambda}E\|Y_{n+1}\|_2^2 + E\|Y_{n+1}\|_2^2\right) \\ &\le \frac{C}{\sqrt{\lambda}}\frac{E\|Y_{n+1}\|_2^2}{1+n}\left(\frac{1+n}{\lambda}+1\right) \le \frac{C}{\sqrt{\lambda}}D\left(\frac{1+n}{\lambda}+1\right) \end{aligned}$$

So when $\lambda/n^{2/3} \to \infty$ we have $\frac{1}{n+1}E_{D_{n+1}}\sum_{i=1}^{n+1}C_3^{[i]} \to 0$.

This completes the proof of Theorem 1. □

Proof of Lemma 1

Proof. We observe that the difference of the first derivatives for the function ϕ_ω satisfies

$$|\phi'_\omega(a) - \phi'_\omega(b)| = \begin{cases} 2(1-\omega)|a-b| & \text{if} \quad (a \le 0, b \le 0), \\ 2\omega|a-b| & \text{if} \quad (a > 0, b > 0), \\ 2|(1-\omega)a - \omega b| & \text{if} \quad (a \le 0, b > 0), \\ 2|\omega a - (1-\omega)b| & \text{if} \quad (a > 0, b \le 0). \end{cases}$$

Therefore we have

$$|\phi'_\omega(a) - \phi'_\omega(b)| \le L|a-b| \quad \forall a, b, \tag{63}$$

where $L = 2\max(1-\omega, \omega)$. By the Lipschitz continuity of ϕ'_ω and Cauchy-Schwarz inequality,

$$(\phi'_\omega(a) - \phi'_\omega(b))(a-b) \le L|a-b|^2 \qquad \forall a, b \in \mathbb{R}. \tag{64}$$

If we let $\varphi_\omega(a) = (L/2)a^2 - \phi_\omega(a)$, then (64) implies the monotonicity of the gradient $\varphi'_\omega(a) = La - \phi'_\omega(a)$. Therefore φ is a convex function and by the first order condition for convexity of φ_ω:

$$\varphi_\omega(a) \ge \varphi_\omega(b) + \varphi'_\omega(b)(a-b) \qquad \forall a, b \in \mathbb{R},$$

which is equivalent to (18). □

Proof of Lemma 2

Proof. 1. By the definition of the majorization function and the fact that $\boldsymbol{\alpha}^{(k+1)}$ is the minimizer in (16)

$$F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k+1)}) \le Q(\boldsymbol{\alpha}^{(k+1)} \mid \boldsymbol{\alpha}^{(k)}) \le Q(\boldsymbol{\alpha}^{(k)} \mid \boldsymbol{\alpha}^{(k)}) = F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)}).$$

2. Based on (20) and the fact that Q is continuous, bounded below and strictly convex, we have

$$\mathbf{0} = \nabla Q(\boldsymbol{\alpha}^{(k+1)} \mid \boldsymbol{\alpha}^{(k)}) = \nabla F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)}) + 2\mathbf{K}_u(\boldsymbol{\alpha}^{(k+1)} - \boldsymbol{\alpha}^{(k)}). \tag{65}$$

Hence

$$
\begin{aligned}
F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k+1)}) &\le Q(\boldsymbol{\alpha}^{(k+1)} \mid \boldsymbol{\alpha}^{(k)}) \\
&= F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)}) + \nabla F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)})(\boldsymbol{\alpha}^{(k+1)} - \boldsymbol{\alpha}^{(k)}) + (\boldsymbol{\alpha}^{(k+1)} - \boldsymbol{\alpha}^{(k)})^\intercal \mathbf{K}_u(\boldsymbol{\alpha}^{(k+1)} - \boldsymbol{\alpha}^{(k)}) \\
&= F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)}) - (\boldsymbol{\alpha}^{(k+1)} - \boldsymbol{\alpha}^{(k)})^\intercal \mathbf{K}_u(\boldsymbol{\alpha}^{(k+1)} - \boldsymbol{\alpha}^{(k)}).
\end{aligned}
$$

By (21) and the assumption that $\sum_{i=1}^{n} \mathbf{K}_i\mathbf{K}_i^\intercal$ is positive definite, we see that $\mathbf{K}_u$ is also positive definite. Let $\gamma_{\min}(\mathbf{K}_u)$ be the smallest eigenvalue of $\mathbf{K}_u$ then

$$
0 \le \gamma_{\min}(\mathbf{K}_u)\|\boldsymbol{\alpha}^{(k+1)}-\boldsymbol{\alpha}^{(k)}\|^2 \le (\boldsymbol{\alpha}^{(k+1)}-\boldsymbol{\alpha}^{(k)})^\intercal \mathbf{K}_u(\boldsymbol{\alpha}^{(k+1)}-\boldsymbol{\alpha}^{(k)}) \le F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)})-F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k+1)}). \tag{66}
$$

Since F is bounded below and monotonically decreasing as shown in Proof 1, $F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)}) - F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k+1)})$ converges to zero as $k \to \infty$, from (66) we see that $\lim_{k\to\infty} \|\boldsymbol{\alpha}^{(k+1)} - \boldsymbol{\alpha}^{(k)}\| = 0$.

3. Now we show that the sequence $(\boldsymbol{\alpha}^{(k)})$ converges to the unique global minimum of (12). As shown in Proof 1, the sequence $(F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)}))$ is monotonically decreasing, hence is bounded above. The fact that $(F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)}))$ is bounded implies that $(\boldsymbol{\alpha}^{(k)})$ must also be bounded, that is because $\lim_{\boldsymbol{\alpha}\to\infty} F_{\omega,\lambda}(\boldsymbol{\alpha}) = \infty$. We next show that the limit of any convergent subsequence of $(\boldsymbol{\alpha}^{(k)})$ is a stationary point of F. Let $(\boldsymbol{\alpha}^{(k_i)})$ be the subsequence of $(\boldsymbol{\alpha}^{(k)})$ and let $\lim_{i\to\infty} \boldsymbol{\alpha}^{(k_i)} = \widehat{\boldsymbol{\alpha}}$, then by (65)

$$
\mathbf{0} = \nabla Q(\boldsymbol{\alpha}^{(k_i+1)} \mid \boldsymbol{\alpha}^{(k_i)}) = \nabla F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k_i)}) + 2\mathbf{K}_u(\boldsymbol{\alpha}^{(k_i+1)} - \boldsymbol{\alpha}^{(k_i)}).
$$

Taking limits on both sides, we prove that $\widehat{\boldsymbol{\alpha}}$ is a stationary point of F.

$$
\begin{aligned}
\mathbf{0} &= \lim_{i\to\infty} \nabla Q(\boldsymbol{\alpha}^{(k_i+1)} \mid \boldsymbol{\alpha}^{(k_i)}) = \nabla Q(\lim_{i\to\infty} \boldsymbol{\alpha}^{(k_i+1)} \mid \lim_{i\to\infty} \boldsymbol{\alpha}^{(k_i)}). \\
&= \nabla F_{\omega,\lambda}(\widehat{\boldsymbol{\alpha}}) + 2\mathbf{K}_u(\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}) = \nabla F_{\omega,\lambda}(\widehat{\boldsymbol{\alpha}}).
\end{aligned}
$$

Then by the strict convexity of F, we have that $\widehat{\boldsymbol{\alpha}}$ is the unique global minimum of (12). □

Proof of Theorem 2

Proof. 1. By (14) and (16),

$$F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k+1)}) \leq Q(\boldsymbol{\alpha}^{(k+1)} \mid \boldsymbol{\alpha}^{(k)}) \leq Q(\Lambda_k \boldsymbol{\alpha}^{(k)} + (1-\Lambda_k)\widehat{\boldsymbol{\alpha}} \mid \boldsymbol{\alpha}^{(k)}). \tag{67}$$

Using (24) we can show that

$$\begin{aligned}
&Q(\Lambda_k \boldsymbol{\alpha}^{(k)} + (1-\Lambda_k)\widehat{\boldsymbol{\alpha}} \mid \boldsymbol{\alpha}^{(k)}) \\
=&F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)}) + (1-\Lambda_k)\nabla F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)})(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(k)}) + (1-\Lambda_k)^2(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(k)})^\intercal \mathbf{K}_u (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(k)}) \\
=&\Lambda_k F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)}) + (1-\Lambda_k)\left[Q(\widehat{\boldsymbol{\alpha}} \mid \boldsymbol{\alpha}^{(k)}) - \Lambda_k(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(k)})^\intercal \mathbf{K}_u (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(k)})\right] \\
=&\Lambda_k F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)}) + (1-\Lambda_k) F_{\omega,\lambda}(\widehat{\boldsymbol{\alpha}}).
\end{aligned} \tag{68}$$

Then the statement can be proved by substituting (68) into (67).

2. We obtain a lower bound for $F_{\omega,\lambda}(\widehat{\boldsymbol{\alpha}})$

$$F_{\omega,\lambda}(\widehat{\boldsymbol{\alpha}}) \geq F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)}) + \nabla F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)})(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(k)}) + (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(k)})^\intercal \mathbf{K}_l (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(k)}), \tag{69}$$

and majorization $Q(\widehat{\boldsymbol{\alpha}} \mid \boldsymbol{\alpha}^{(k)})$

$$Q(\widehat{\boldsymbol{\alpha}} \mid \boldsymbol{\alpha}^{(k)}) = F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)}) + \nabla F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)})(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(k)}) + (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(k)})^\intercal \mathbf{K}_u (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(k)}). \tag{70}$$

Subtract (69) from (70) and divide by $(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(k)})^\intercal \mathbf{K}_u (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(k)})$, we have

$$\begin{aligned}
\Lambda_k &= \frac{Q(\widehat{\boldsymbol{\alpha}} \mid \boldsymbol{\alpha}^{(k)}) - F_{\omega,\lambda}(\widehat{\boldsymbol{\alpha}})}{(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(k)})^\intercal \mathbf{K}_u (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(k)})} \\
&\leq 1 - \frac{(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(k)})^\intercal \mathbf{K}_l (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(k)})}{(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(k)})^\intercal \mathbf{K}_u (\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(k)})} \\
&\leq 1 - \gamma_{\min}(\mathbf{K}_u^{-1}\mathbf{K}_l).
\end{aligned} \tag{71}$$

Both K_u and K_l are positive definite by the assumption that $\sum_{i=1}^n \mathbf{K}_i \mathbf{K}_i^\intercal$ is positive definite, and since

$$\mathbf{K}_u^{-1}\mathbf{K}_l = \mathbf{K}_u^{-\frac{1}{2}} \mathbf{K}_u^{-\frac{1}{2}} \mathbf{K}_l \mathbf{K}_u^{-\frac{1}{2}} \mathbf{K}_u^{\frac{1}{2}},$$

the matrix $\mathbf{K}_u^{-1}\mathbf{K}_l$ is similar to the matrix $\mathbf{K}_u^{-\frac{1}{2}}\mathbf{K}_l\mathbf{K}_u^{-\frac{1}{2}}$, which is positive definite. Hence

$$\Gamma = 1 - \gamma_{\min}(\mathbf{K}_u^{-1}\mathbf{K}_l) = 1 - \gamma_{\min}(\mathbf{K}_u^{-\frac{1}{2}}\mathbf{K}_l\mathbf{K}_u^{-\frac{1}{2}}) < 1.$$

By (14) and (71) we showed that $0 \leq \Lambda_k \leq \Gamma < 1$.

3. Since $\nabla F_{\omega,\lambda}(\widehat{\boldsymbol{\alpha}}) = \mathbf{0}$, using the Taylor expansion on $F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)})$ at $\widehat{\boldsymbol{\alpha}}$, we have

$$F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)}) - F_{\omega,\lambda}(\widehat{\boldsymbol{\alpha}}) \geq (\boldsymbol{\alpha}^{(k)} - \widehat{\boldsymbol{\alpha}})^\intercal\mathbf{K}_l(\boldsymbol{\alpha}^{(k)} - \widehat{\boldsymbol{\alpha}}) \geq \gamma_{\min}(\mathbf{K}_l)\|\boldsymbol{\alpha}^{(k)} - \widehat{\boldsymbol{\alpha}}\|^2,$$

$$F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)}) - F_{\omega,\lambda}(\widehat{\boldsymbol{\alpha}}) \leq (\boldsymbol{\alpha}^{(k)} - \widehat{\boldsymbol{\alpha}})^\intercal\mathbf{K}_u(\boldsymbol{\alpha}^{(k)} - \widehat{\boldsymbol{\alpha}}) \leq \gamma_{\max}(\mathbf{K}_u)\|\boldsymbol{\alpha}^{(k)} - \widehat{\boldsymbol{\alpha}}\|^2.$$

Therefore, by Results 1 and 2

$$\|\boldsymbol{\alpha}^{(k+1)}-\widehat{\boldsymbol{\alpha}}\|^2 \leq \frac{F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k+1)}) - F_{\omega,\lambda}(\widehat{\boldsymbol{\alpha}})}{\gamma_{\min}(\mathbf{K}_l)} \leq \frac{\Gamma(F_{\omega,\lambda}(\boldsymbol{\alpha}^{(k)}) - F_{\omega,\lambda}(\widehat{\boldsymbol{\alpha}}))}{\gamma_{\min}(\mathbf{K}_l)} \leq \Gamma\frac{\gamma_{\max}(\mathbf{K}_u)}{\gamma_{\min}(\mathbf{K}_l)}\|\boldsymbol{\alpha}^{(k)}-\widehat{\boldsymbol{\alpha}}\|^2.$$

□