Supplementary materials for the manuscript "A Scalable Algorithm for Structured Kernel Feature Selection"

1 Convergence and Regret Analysis and Proof for Theorem 2

We provide the detailed proof for **Theorem 2** in **Section 3.2** in the manuscript. First, we re-state the theorem as follows:

Theorem 2 With an auxiliary function $h(\mathbf{a}) = ||\mathbf{a}||^2$, and the non-decreasing sequence $\{\beta_t\}$ with $\beta_t = \gamma(1+\ln(t))$, Let $\{\mathbf{a}_t\}$ and $\{\mathbf{g}_t\}$ be two sequences generated by **Algorithm 1** in the manuscript. Suppose the optimal solution $\mathbf{a}^*$ to the original problem (1) in the manuscript satisfies $h(\mathbf{a}^*) \leq D$, for some $D > 0$, and there is a constant G such that $||\mathbf{g}_t||_* \leq G$ for all $t \geq 1$, we have the following properties for **Algorithm 1**:
a) For each $t \geq 1$, the average regret is bounded by

$$R_t(\mathbf{a}) \leq \left(\gamma D^2 + \frac{G^2}{2\gamma}\right)(1+\ln(t)).$$

b) The sequence of primal variables are bounded by

$$||\mathbf{a}_{t+1} - \mathbf{a}^*|| \leq \frac{2}{\gamma(1+t+\ln(t))}\left(\left(\gamma D^2 + \frac{G^2}{2\gamma}\right)(1+\ln(t)) - R_t(\mathbf{a}^*)\right).$$

Also we can have the convergence in the expectation form:
c)

$$\mathbf{E}||\mathbf{a}_{t+1} - \mathbf{a}^*|| \leq \frac{2}{1+t+\ln(t)}\left(D^2 + \frac{G^2}{2\gamma^2}\right)(1+\ln(t)).$$

Proof: We use the indication function to represent the non-negative region constraint:

$$\Phi(\mathbf{a}) = I_C(\mathbf{a}) = \begin{cases} 0 & \text{if } a_i \geq 0, \forall i > 0 \\ \infty & \text{if } \exists a_i < 0, i > 0 \end{cases}$$

The loss function for our original problem can be written as:

$$f(\mathbf{a}) = \sum_{m=1}^{n}[n - \bar{L}_m^T(a_0\mathbf{1} + \sum_{i=1}^{p} a_i \bar{K}_m^i)]_+ + \lambda_1 \sum_{i=1}^{p} a_i + \lambda_2 \sum_{(i,j)\in E}(a_i - a_j)^2$$

We define the region

$$\mathscr{F}_D = \{\mathbf{a} \in \mathrm{dom}(\Phi) | h(\mathbf{a}) \leq D^2\}.$$

a) For the regret analysis, let

$$\delta_t = \max_{\mathbf{a}\in\mathscr{F}_D}\left\{\sum_{\zeta=1}^{t}\big(\langle \mathbf{g}_\zeta, \mathbf{a}_\zeta - \mathbf{a}\rangle + \Phi(\mathbf{a}_\zeta)\big) - t\Phi(\mathbf{a}))\right\}, \quad t = 1, 2, 3, ...$$

We can see that δ_t is the upper bound of the regret $R_t(\mathbf{a})$

$$\begin{aligned} R_t(\mathbf{a}) &= \sum_{\zeta=1}^{t}\big(f_\zeta(\mathbf{a}_\zeta) + \Phi(\mathbf{a}_\zeta)\big) - \sum_{\zeta=1}^{t}\big(f_\zeta(\mathbf{a}) + \Phi(\mathbf{a})\big) \\ &= \sum_{\zeta=1}^{t}\big(f_\zeta(\mathbf{a}_\zeta) - f_\zeta(\mathbf{a}) + \Phi(\mathbf{a}_\zeta)\big) - t\Phi(\mathbf{a}) \\ &\leq \sum_{\zeta=1}^{t}\big(\langle \mathbf{g}_\zeta, \mathbf{a}_\zeta - \mathbf{a}\rangle + \Phi(\mathbf{a}_\zeta)\big) - t\Phi(\mathbf{a}) \\ &\leq \delta_t \end{aligned} \tag{1}$$

For an arbitrary initial feasible solution $\mathbf{a}_0$, we can rewrite

$$\begin{aligned} \delta_t = &\sum_{\zeta=1}^{t}\big(\langle \mathbf{g}_\zeta, \mathbf{a}_\zeta - \mathbf{a}_0\rangle + \Phi(\mathbf{a}_\zeta)\big) \\ &+ \max_{\mathbf{a}\in\mathscr{F}_D}\big\{\langle t\bar{\mathbf{g}}_t, \mathbf{a}_0 - \mathbf{a}\rangle - t\Phi(\mathbf{a})\big\}. \end{aligned}$$

Define $V_t(t\bar{\mathbf{g}}_t) = \max_{\mathbf{a}}\big\{\langle t\bar{\mathbf{g}}_t, \mathbf{a} - \mathbf{a}_0\rangle - t\Phi(\mathbf{a}) - \beta_t h(\mathbf{a})\big\}$. As $\mathbf{a} \in \mathscr{F}_D$, we can derive the following inequality similarly as in Lemma 9 in (Xiao, 2010):

$$\delta_t \leq \sum_{\zeta=1}^{t}\big(\langle \mathbf{g}_\zeta, \mathbf{a}_\zeta - \mathbf{a}_0\rangle + \Phi(\mathbf{a}_\zeta)\big) + V_t(-t\bar{\mathbf{g}}_t) + \beta_t D^2. \tag{2}$$

According to Lemmas 10 and 11 in (Xiao, 2010), we can easily get

$$V_\zeta(-\zeta\bar{\mathbf{g}}_\zeta) + \Phi(\mathbf{a}_{\zeta+1}) \leq V_\zeta(-\zeta\bar{\mathbf{g}}_\zeta),$$

and

$$V_\zeta(-\zeta\bar{\mathbf{g}}_\zeta) \le V_{\zeta-1}(-(\zeta-1)\bar{\mathbf{g}}_{\zeta-1}) + \langle -\mathbf{g}_\zeta, \mathbf{a}_\zeta - \mathbf{a}_0\rangle + \frac{||\mathbf{g}_\zeta||_*^2}{2(\gamma(\zeta-1)+\beta_{\zeta-1})}$$

when $\zeta \ge 2$. Hence

$$V_\zeta(-\zeta\bar{\mathbf{g}}_\zeta) + \Phi(\mathbf{a}_{\zeta+1}) \le V_{\zeta-1}(-(\zeta-1)\bar{\mathbf{g}}_{\zeta-1}) + \langle -\mathbf{g}_\zeta, \mathbf{a}_\zeta - \mathbf{a}_0\rangle + \frac{||\mathbf{g}_\zeta||_*^2}{2(\gamma(\zeta-1)+\beta_{\zeta-1})}, \zeta \ge 2.$$

Moving corresponding terms, we get:

$$\langle \mathbf{g}_\zeta, \mathbf{a}_\zeta - \mathbf{a}_0\rangle + \Phi(\mathbf{a}_{\zeta+1}) \le V_{\zeta-1}(-(\zeta-1)\bar{\mathbf{g}}_{\zeta-1}) - V_\zeta(-\zeta\bar{\mathbf{g}}_\zeta) + \frac{||\mathbf{g}_\zeta||_*^2}{2(\gamma(\zeta-1)+\beta_{\zeta-1})}, \zeta \ge 2.$$

When $\zeta = 1$, we have

$$\langle \mathbf{g}_1, \mathbf{a}_1 - \mathbf{a}_0\rangle + \Phi(\mathbf{a}_2) \le - V_1(-\bar{\mathbf{g}}_1) + \frac{||\mathbf{g}_1||_*^2}{2(\beta_0)} + (\beta_0 - \beta_1)h(\mathbf{a}_2)$$

By adding all the inequalities for $\zeta = 1, ..., t$, we can get

$$\sum_{\zeta=1}^{t} \big(\langle \mathbf{g}_\zeta, \mathbf{a}_\zeta - \mathbf{a}_0\rangle + \Phi(\mathbf{a}_{\zeta+1})\big) + V_\zeta(-\zeta\bar{\mathbf{g}}_\zeta) \le (\beta_0 - \beta_1)h(\mathbf{a}_2) + \frac{1}{2}\sum_{\zeta=1}^{t} \frac{||g_\zeta||_*^2}{\gamma(\zeta-1)+\beta_{\zeta-1}}$$

Since $\mathbf{a}_1 = \mathbf{a}_0 = \mathbf{0} \in argmin_{\mathbf{a}}\Phi(\mathbf{a})$, so $\Phi(\mathbf{a}_{t+1}) \ge \Phi(\mathbf{a}_0) = \Phi(\mathbf{a}_1)$. Adding $\Phi(\mathbf{a}_1) - \Phi(\mathbf{a}_{t+1})$ to both sides,

$$\sum_{\zeta=1}^{t} \big(\langle \mathbf{g}_\zeta, \mathbf{a}_\zeta - \mathbf{a}_0\rangle + \Phi(\mathbf{a}_\zeta)\big) + V_\zeta(-\zeta\bar{\mathbf{g}}_\zeta) \tag{3}$$
$$\le (\beta_0 - \beta_1)h(\mathbf{a}_2) + \tfrac{1}{2}\sum_{\zeta=1}^{t} \tfrac{||g_\zeta||_*^2}{\gamma(\zeta-1)+\beta_{\zeta-1}} \tag{4}$$

Substituting this into (2) , we have

$$R_t(\mathbf{a}) \le \delta_t \le \beta_t D^2 + \frac{1}{2}\sum_{\zeta=1}^{t} \frac{||g_\zeta||_*^2}{\gamma(\zeta-1)+\beta_{\zeta-1}} + \frac{2(\beta_0-\beta_1)||\mathbf{g}_1||_*^2}{\beta_1+\gamma}.$$

For our algorithm $\beta_t = \gamma(1+\ln(t))$, and $\beta_0 = \beta_1 = \gamma$, hence

$$R_t(\mathbf{a}) \le \delta_t \le \gamma(1+\ln(t))D^2 + \frac{G^2}{2\gamma}\big(1 + \sum_{\zeta=1}^{t-1} \frac{1}{\zeta+1+\ln\zeta}\big) \le \left(\gamma D^2 + \frac{G^2}{2\gamma}\right)(1+\ln(t))$$

b) To find the bounds for primal variables, we first rewrite the solution to the subproblem (9) in the manuscript at the tth step in **Algorithm 1**:

$$\mathbf{a}_{t+1} = \arg\min_{\mathbf{a}} \big\{\langle t\bar{\mathbf{g}}_t, \mathbf{a}\rangle + t\Phi(\mathbf{a}) + \beta_t h(\mathbf{a})\big\}.$$

The subgradients $\mathbf{b}_{t+1} \in \partial\Phi(\mathbf{a}_{t+1})$ and $\mathbf{d}_{t+1} \in \partial h(\mathbf{a}_{t+1})$ satisfy the following inequality:

$$\langle t\bar{\mathbf{g}}_t + t\mathbf{b}_{t+1} + \beta_t\mathbf{d}_{t+1}, \mathbf{a} - \mathbf{a}_{t+1}\rangle \ge 0, \forall \mathbf{a} \in \mathrm{dom}(\Phi).$$

Since both $\Phi(\cdot)$ and $h(\cdot)$ are strongly convex, we have

$$\begin{aligned}
&\frac{1}{2}(\gamma t + \beta_t)||\mathbf{a}_{t+1} - \mathbf{a}||^2 \\
&\le t\big(\Phi(\mathbf{a}) - \Phi(\mathbf{a}_{t+1}) - \langle \mathbf{b}_{t+1}, \mathbf{a} - \mathbf{a}_{t+1}\rangle\big) + \\
&\quad \beta_t\big(h(\mathbf{a}) - h(\mathbf{a}_{t+1}) - \langle \mathbf{d}_{t+1}, \mathbf{a} - \mathbf{a}_{t+1}\rangle\big) \\
&= \beta_t h(\mathbf{a}) - \beta_t h(\mathbf{a}_{t+1}) - \langle t\mathbf{b}_{t+1} + \beta_t\mathbf{d}_{t+1}, \mathbf{a} - \mathbf{a}_{t+1}\rangle \\
&\quad + t\Phi(\mathbf{a}) - t\Phi(\mathbf{a}_{t+1}) \\
&\le \beta_t h(\mathbf{a}) - \beta_t h(\mathbf{a}_{t+1}) + \langle t\bar{\mathbf{g}}_t, \mathbf{a} - \mathbf{a}_{t+1}\rangle + t\Phi(\mathbf{a}) \\
&\quad - t\Phi(\mathbf{a}_{t+1}) \\
&= \beta_t h(\mathbf{a}) + t\Phi(\mathbf{a}) + \big\{\langle -t\bar{\mathbf{g}}_t, \mathbf{a}_{t+1} - \mathbf{a}_0\rangle - \beta_t h(\mathbf{a}_{t+1}) \\
&\quad - t\Phi(\mathbf{a}_{t+1})\big\} + \langle t\bar{\mathbf{g}}_t, \mathbf{a} - \mathbf{a}_0\rangle \\
&= \beta_t h(\mathbf{a}) + t\Phi(\mathbf{a}) + V_t(-t\bar{\mathbf{g}}_t) + \langle t\bar{\mathbf{g}}_t, \mathbf{a} - \mathbf{a}_0\rangle.
\end{aligned}$$

Note that for the dual average methods in **Algorithm 1**,

$$\langle t\bar{\mathbf{g}}_t, \mathbf{a} - \mathbf{a}_0\rangle = \sum_{\zeta=1}^{t}\langle \mathbf{g}_\zeta, \mathbf{a} - \mathbf{a}_\zeta\rangle + \sum_{\zeta=1}^{t}\langle \mathbf{g}_\zeta, \mathbf{a}_\zeta - \mathbf{a}_0\rangle.$$

Substituting the corresponding term, we can get

$$\begin{aligned}
&\frac{1}{2}(\gamma t + \beta_t)||\mathbf{a}_{t+1} - \mathbf{a}||^2 \\
&\le \beta_t h(\mathbf{a}) + \left\{V_t(-t\bar{\mathbf{g}}_t) + \sum_{\zeta=1}^{t}\big(\langle \mathbf{g}_\zeta, \mathbf{a} - \mathbf{a}_0\rangle + \Phi(\mathbf{a}_\zeta)\big)\right\} \\
&+ \sum_{\zeta=1}^{t}\langle \mathbf{g}_\zeta, \mathbf{a} - \mathbf{a}_\zeta\rangle + t\Phi(\mathbf{a}) - \sum_{\zeta=1}^{t}\Phi(\mathbf{a}_\zeta).
\end{aligned}$$

Taking the proof for a) (1) that

$$\begin{aligned}
&\sum_{\zeta=1}^{t}\langle \mathbf{g}_\zeta, \mathbf{a} - \mathbf{a}_\zeta\rangle + t\Phi(\mathbf{a}) - \sum_{\zeta=1}^{t}\Phi(\mathbf{a}_\zeta) \\
&\le \sum_{\zeta=1}^{t}\big(f_\zeta(\mathbf{a}) - f_\zeta(\mathbf{a}_\zeta)\big) + t\Phi(\mathbf{a}) - \sum_{\zeta=1}^{t}\Phi(\mathbf{a}_\zeta) \\
&= \sum_{\zeta=1}^{t}\big(f_\zeta(\mathbf{a}) + \Phi(\mathbf{a})\big) - \sum_{\zeta=1}^{t}\big(f_\zeta(\mathbf{a}_\zeta) + \Phi(\mathbf{a}_\zeta)\big) \\
&= -R_t(\mathbf{a}),
\end{aligned}$$

Using (4), we can derive

$$\frac{1}{2}(\gamma t + \beta_t)||\mathbf{a}_{t+1} - \mathbf{a}||_2^2 \le \beta_t h(\mathbf{a}) + (\beta_0 - \beta_1)h(\mathbf{a}_2) + \frac{1}{2}\sum_{\zeta=1}^{t} \frac{||\mathbf{g}_\zeta||_*^2}{\gamma(\zeta-1)+\beta_{\zeta-1}} - R_t(\mathbf{a})$$

By the assumptions given in the theorem, and setting $\beta_0 = \beta_1 = \gamma$, we have

$$\frac{1}{2}(\gamma t + \beta_t)||\mathbf{a}_{t+1} - \mathbf{a}||_2^2 \le$$

$$\gamma(1+\ln(t))D^2 + \frac{G^2}{2\gamma}\left(1 + \sum_{\zeta=1}^{t-1}\frac{1}{\zeta+1+\ln\zeta}\right) - R_t(\mathbf{a})$$

$$\le \left(\gamma D^2 + \frac{G^2}{2\gamma}\right)(1+\ln(t)) - R_t(\mathbf{a}).$$

Hence,

$$||\mathbf{a}_{t+1} - \mathbf{a}^*|| \le \frac{2}{\gamma(1+t+\ln(t))}\left(\left(\gamma D^2 + \frac{G^2}{2\gamma}\right)(1+\ln(t)) - R_t(\mathbf{a}^*)\right).$$

c) Let $z_\zeta = \{Y_\zeta, X_\zeta\}$ be the ζth sample for **Algorithm 1**, and $\mathbf{z}[t]$ denote the collection of i.i.d random variables $\{z_1, ..., z_t\}$. We can take $\mathbf{a}_\zeta$ as a function of $\{z_1, ..., z_{\zeta-1}\}$, which is independent of $\{z_\zeta, ..., z_t\}$.

We have

$$R_t(\mathbf{a}^*) = \sum_{\zeta=1}^{t}\big(f(\mathbf{a}_\zeta, z_\zeta) + \Phi(\mathbf{a}_\zeta)\big) - \sum_{\zeta=1}^{t}\big(f(\mathbf{a}_\zeta^*, z_\zeta) + \Phi(\mathbf{a}_\zeta^*)\big),$$

and

$$\begin{aligned}\mathbf{E}_{\mathbf{z}[t]}\big(f(\mathbf{a}_\zeta, z_\zeta) + \Phi(\mathbf{a}_\zeta)\big) &= \mathbf{E}_{\mathbf{z}[\zeta-1]}\big(f(\mathbf{a}_\zeta, z_\zeta) + \Phi(\mathbf{a}_\zeta)\big)\\ &= \mathbf{E}_{\mathbf{z}[t]}\big(f(\mathbf{a}_\zeta) + \Phi(\mathbf{a}_\zeta)\big).\end{aligned}$$

We also can get

$$\begin{aligned}\mathbf{E}_{\mathbf{z}[t]}\big(f(\mathbf{a}^*, z_\zeta) + \Phi(\mathbf{a}^*)\big) &= \mathbf{E}_{z_\zeta}\big(f(\mathbf{a}^*, z_\zeta) + \Phi(\mathbf{a}^*)\big)\\ &= f(\mathbf{a}^*) + \Phi(\mathbf{a}^*).\end{aligned}$$

Since

$$f(\mathbf{a}^*) + \Phi(\mathbf{a}^*) = \min_{\mathbf{a}} f(\mathbf{a}) + \Phi(\mathbf{a}),$$

combining the previous results leads to the following equation:

$$\begin{aligned}\mathbf{E}_{\mathbf{z}[t]}R_t(\mathbf{a}^*) &= \sum_{\zeta=1}^{t}\mathbf{E}_{\mathbf{z}[t]}\big(f(\mathbf{a}_\zeta) + \Phi(\mathbf{a}_\zeta)\big) - t\big(f(\mathbf{a}^*) + \Phi(\mathbf{a}^*)\big)\\ &\ge 0.\end{aligned}$$

Therefore, with the result from b), we can get

$$\mathbf{E}||\mathbf{a}_{t+1} - \mathbf{a}^*|| \le \frac{2}{1+t+\ln(t)}\left(D^2 + \frac{G^2}{2\gamma^2}\right)(1+\ln(t)).$$

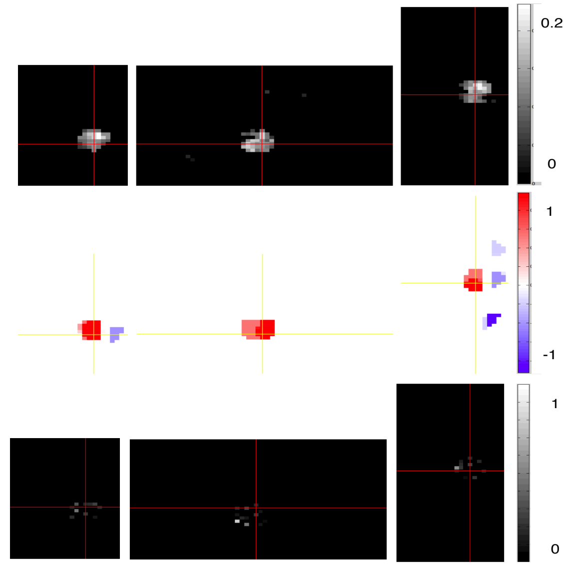

Figure 1: Active Regions recovered by the proposed method, Fused LASSO and HSIC-LASSO for simulated MRI images with additive nonlinear responses.

2 Some Figures for Simulation Results

Figures 1 and 2 illustrate the results for additive nonlinear and non-additive nonlinear simulations in **Sections 4.1.2** and **4.1.3** in the manuscript.

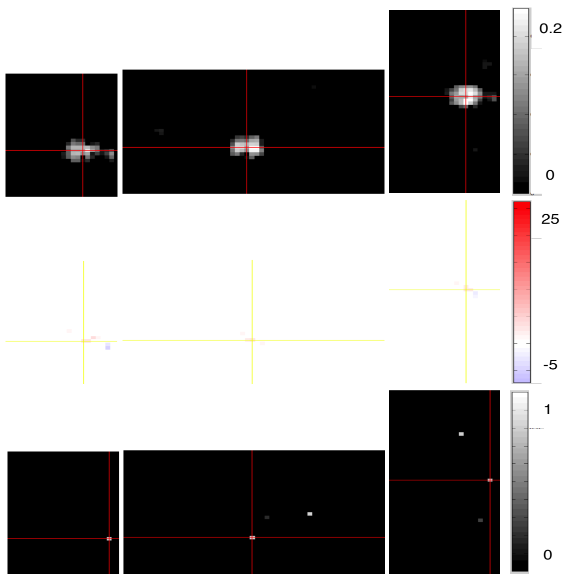

Figure 2: Active Regions recovered by the proposed method, Fused LASSO and HSIC-LASSO for simulated MRI images with non-additive nonlinear responses.