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 \( h(a) = ||a||^2 \), and the non-decreasing sequence \( \{\beta_t\} \) with \( \beta_t = \gamma_t(1 + \ln(t)) \), Let \( \{a_t\} \) and \( \{g_t\} \) be two sequences generated by Algorithm 1 in the manuscript. Suppose the optimal solution \( a^* \) to the original problem (1) in the manuscript satisfies \( h(a^*) \leq D \), for some \( D > 0 \), and there is a constant \( G \) such that \( ||g_t||_2 \leq G \) 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

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

b) The sequence of primal variables are bounded by

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

We define the region

\[
\mathcal{F}_D = \{a \in \text{dom}(\Phi) | h(a) \leq D^2\}.
\]

a) For the regret analysis, let

\[
\delta_t = \max_{a \in \mathcal{F}_D} \left\{ \sum_{\zeta=1}^{t} \left( (g_\zeta, a_\zeta - a) + \Phi(a_\zeta) \right) - t\Phi(a) \right\},
\]

\[
t = 1, 2, 3, ...
\]

We can see that \( \delta_t \) is the upper bound of the regret \( R_t(a) \)

\[
R_t(a) = \sum_{\zeta=1}^{t} \left( f_\zeta(a_\zeta) - f_\zeta(a) + \Phi(a_\zeta) \right) - \sum_{\zeta=1}^{t} \left( f_\zeta(a) + \Phi(a) \right)
\]

\[
= \sum_{\zeta=1}^{t} \left( f_\zeta(a_\zeta) - f_\zeta(a) + \Phi(a_\zeta) \right) - t\Phi(a)
\]

\[
\leq \sum_{\zeta=1}^{t} \left( (g_\zeta, a_\zeta - a) + \Phi(a_\zeta) \right) - t\Phi(a)
\]

\[
\leq \delta_t \tag{1}
\]

For an arbitrary initial feasible solution \( a_0 \), we can rewrite

\[
\delta_t = \sum_{\zeta=1}^{t} \left( (g_\zeta, a_\zeta - a_0) + \Phi(a_\zeta) \right)
\]

\[
+ \max_{a \in \mathcal{F}_D} \left\{ (t\bar{g}_t, a_0 - a) - t\Phi(a) \right\}.
\]

Define \( V_t(t\bar{g}_t) = \max_{a} \left\{ (t\bar{g}_t, a - a_0) - t\Phi(a) - \beta_i h(a) \right\} \). As \( a \in \mathcal{F}_D \), we can derive the following inequality similarly as in Lemma 9 in (Xiao, 2010):

\[
\delta_t \leq \sum_{\zeta=1}^{t} \left( (g_\zeta, a_\zeta - a_0) + \Phi(a_\zeta) \right) + V_{t}(-t\bar{g}_t) + \beta_i D^2. \tag{2}
\]

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

\[
V_{\zeta}(-\zeta\bar{g}_\zeta) + \Phi(a_{\zeta+1}) \leq V_{\zeta}(-\zeta\bar{g}_\zeta),
\]

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

\[
f(a) = \sum_{m=1}^{n} \left[ n - L^m(a_0) + \sum_{i=1}^{p} a_i R^m_i \right] + \lambda_1 \sum_{i=1}^{p} a_i
\]

\[
+ \lambda_2 \sum_{(i, j) \in E} (a_i - a_j)^2
\]

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

The supplementary materials for the manuscript “A Scalable Algorithm for Structured Kernel Feature Selection.”
and
\[ V_{\zeta}(-\zeta \bar{g}_{\zeta}) \leq V_{\zeta-1}(-\zeta + 1) + \langle -g_{\zeta}, a_{\zeta} - a_0 \rangle + \frac{||g_{\zeta}||^2}{2(\gamma(\zeta - 1) + \beta_{\zeta-1})} \]
when \( \zeta \geq 2 \). Hence
\[ V_{\zeta}(-\zeta \bar{g}_{\zeta}) + \Phi(a_{\zeta+1}) \leq V_{\zeta-1}(-\zeta + 1) + \langle -g_{\zeta}, a_{\zeta} - a_0 \rangle + \frac{||g_{\zeta}||^2}{2(\gamma(\zeta - 1) + \beta_{\zeta-1})}, \zeta \geq 2. \]

Moving corresponding terms, we get:
\[ \langle g_{\zeta}, a_{\zeta} - a_0 \rangle + \Phi(a_{\zeta+1}) \leq V_{\zeta-1}(-\zeta + 1) \]
\[ - V_{\zeta}(-\zeta \bar{g}_{\zeta}) + \frac{||g_{\zeta}||^2}{2(\gamma(\zeta - 1) + \beta_{\zeta-1})}, \zeta \geq 2. \]

When \( \zeta = 1 \), we have
\[ \langle g_1, a_1 - a_0 \rangle + \Phi(a_2) \leq -V_1(-\bar{g}_1) + \frac{||g_1||^2}{2(\beta_0)} \]
\[ + (\beta_0 - \beta_1)h(a_2) \]

By adding all the inequalities for \( \zeta = 1, ..., t \), we can get
\[ \sum_{\zeta=1}^t \left( \langle g_{\zeta}, a_{\zeta} - a_0 \rangle + \Phi(a_{\zeta+1}) \right) + V_{\zeta}(-\zeta \bar{g}_{\zeta}) \]
\[ \leq (\beta_0 - \beta_1)h(a_2) + \frac{1}{2} \sum_{\zeta=1}^t \frac{||g_{\zeta}||^2}{\gamma(\zeta - 1) + \beta_{\zeta-1}} \]

Since \( a_1 = a_0 = 0 \in \text{argmin}_a \Phi(a) \), so \( \Phi(a_{\zeta+1}) \geq \Phi(a_0) = \Phi(a_1) \). Adding \( \Phi(a_1) - \Phi(a_{\zeta+1}) \) to both sides,
\[ \sum_{\zeta=1}^t \left( \langle g_{\zeta}, a_{\zeta} - a_0 \rangle + \Phi(a_{\zeta}) \right) + V_{\zeta}(-\zeta \bar{g}_{\zeta}) \]
\[ \leq (\beta_0 - \beta_1)h(a_2) + \frac{1}{2} \sum_{\zeta=1}^t \frac{||g_{\zeta}||^2}{\gamma(\zeta - 1) + \beta_{\zeta-1}} \]

Substituting this into (2), we have
\[ R_t(a) \leq \delta_t \leq \beta_t D^2 + \frac{1}{2} \sum_{\zeta=1}^t \frac{||g_{\zeta}||^2}{\gamma(\zeta - 1) + \beta_{\zeta-1}} \]
\[ + \frac{2(\beta_0 - \beta_1)||g_1||^2}{\beta_1 + \gamma}. \]

For our algorithm \( \beta_t = \gamma(1 + \ln(t)) \), and \( \beta_0 = \beta_1 = \gamma \), hence
\[ R_t(a) \leq \delta_t \leq \gamma(1 + \ln(t)) D^2 + \frac{G^2}{2\gamma} \left( 1 + \sum_{\zeta=1}^{t-1} \frac{1}{\zeta + 1 + \ln \zeta} \right) \]
\[ \leq \left( \gamma D^2 + \frac{G^2}{2\gamma} \right) (1 + \ln(t)) \]

b) To find the bounds for primal variables, we first rewrite the solution to the subproblem (9) in the manuscript at the \( t \)th step in Algorithm 1:
\[ a_{t+1} = \arg \min_a \{ \langle \bar{g}_{t}, a \rangle + t\Phi(a) + \beta_t h(a) \}. \]

The subgradients \( b_{t+1} \in \partial \Phi(a_{t+1}) \) and \( d_{t+1} \in \partial h(a_{t+1}) \) satisfy the following inequality:
\[ \langle \bar{g}_{t}, t b_{t+1} + \beta_t d_{t+1}, a - a_{t+1} \rangle \geq 0, \forall a \in \text{dom}(\Phi). \]

Since both \( \Phi(\cdot) \) and \( h(\cdot) \) are strongly convex, we have
\[ \frac{1}{2} (\gamma t + \beta_t)||a_{t+1} - a||^2 \]
\[ \leq t \left( \Phi(a) - \Phi(a_{t+1}) - \langle b_{t+1}, a - a_{t+1} \rangle \right) + \beta_t h(a) - h(a_{t+1}) \]
\[ = \beta_t h(a) - \beta_t h(a_{t+1}) - (t b_{t+1} + \beta_t d_{t+1}, a - a_{t+1}) + t \Phi(a) - t \Phi(a_{t+1}) \]
\[ \leq \beta_t h(a) + \Phi(a) + \left( \langle -\bar{g}_{t}, a_{t+1} - a_0 \rangle - \beta_t h(a_{t+1}) \right. \]
\[ \left. - t \Phi(a_{t+1}) \right) + \left( \langle \bar{g}_{t}, a - a_0 \rangle \right) \]
\[ \leq \beta_t h(a) + t \Phi(a) + V_{t}(-t \bar{g}_t) + \langle \bar{g}_{t}, a - a_0 \rangle. \]

Note that for the dual average methods in Algorithm 1,
\[ \langle \bar{g}_{t}, a - a_0 \rangle = \sum_{\zeta=1}^t \langle g_\zeta, a - a_\zeta \rangle + \sum_{\zeta=1}^t \langle g_\zeta, a_\zeta - a_0 \rangle. \]

Substituting the corresponding term, we can get
\[ \frac{1}{2} (\gamma t + \beta_t)||a_{t+1} - a||^2 \]
\[ \leq \beta_t h(a) + \left( V_{t}(-t \bar{g}_t) + \sum_{\zeta=1}^t \left( \langle g_\zeta, a - a_0 \rangle + \Phi(a_\zeta) \right) \right) \]
\[ + \sum_{\zeta=1}^t \langle g_\zeta, a - a_\zeta \rangle + t \Phi(a) - \sum_{\zeta=1}^t \Phi(a_\zeta). \]

Taking the proof for \( a_1 \) (1) that
\[ \sum_{\zeta=1}^t \langle g_\zeta, a - a_\zeta \rangle + t \Phi(a) - \sum_{\zeta=1}^t \Phi(a_\zeta) \]
\[ \leq \sum_{\zeta=1}^t (f_\zeta(a) - f_\zeta(a_\zeta)) + t \Phi(a) - \sum_{\zeta=1}^t \Phi(a_\zeta) \]
\[ = \sum_{\zeta=1}^t (f_\zeta(a) + \Phi(a)) - \sum_{\zeta=1}^t (f_\zeta(a_\zeta) + \Phi(a_\zeta)) \]
\[ = -R_t(a), \]
Using (4), we can derive
\[ \frac{1}{2} (\gamma t + \beta_t) ||a_{t+1} - a||^2 \leq \beta_t h(a) + (\beta_0 - \beta_1) h(a_2) + \frac{1}{2} \sum_{\zeta=1}^{t} \frac{||g_\zeta||^2}{\gamma(\zeta - 1) + \beta_{\zeta-1}} - R_t(a) \]

By the assumptions given in the theorem, and setting \( \beta_0 = \beta_1 = \gamma \), we have
\[ \frac{1}{2} (\gamma t + \beta_t) ||a_{t+1} - a||^2 \leq \gamma(1 + \ln(t))D^2 + \frac{G^2}{2\gamma} \left( 1 + \sum_{\zeta=1}^{t-1} \frac{1}{\zeta + 1 + \ln \zeta} \right) - R_t(a) \]
\[ \leq \left( \gamma D^2 + \frac{G^2}{2\gamma} \right) (1 + \ln(t)) - R_t(a). \]

Hence,
\[ ||a_{t+1} - a^*|| \leq \frac{2}{\gamma(1 + t + \ln(t))} \left( \gamma D^2 + \frac{G^2}{2\gamma} \right) (1 + \ln(t)) - R_t(a^*) \]

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

We have
\[ R_t(a^*) = \sum_{\zeta=1}^{t} (f(a_\zeta, z_\zeta) + \Phi(a_\zeta)) - \sum_{\zeta=1}^{t} (f(a^*_\zeta, z_\zeta) + \Phi(a^*_\zeta)), \]
and
\[ E_{z[t]}(f(a_\zeta, z_\zeta) + \Phi(a_\zeta)) = E_{z[\zeta-1]}(f(a_\zeta, z_\zeta) + \Phi(a_\zeta)) = E_{z[t]}(f(a_\zeta) + \Phi(a_\zeta)). \]

We also can get
\[ E_{z[t]}(f(a^*_\zeta, z_\zeta) + \Phi(a^*)) = E_{z_\zeta}(f(a^*_\zeta, z_\zeta) + \Phi(a^*)) = f(a^*) + \Phi(a^*). \]

Since
\[ f(a^*) + \Phi(a^*) = \min_a f(a) + \Phi(a), \]
combining the previous results leads to the following equation:
\[ E_{z[t]} R_t(a^*) = \sum_{\zeta=1}^{t} E_{z[t]}(f(a_\zeta) + \Phi(a_\zeta)) - t(f(a^*) + \Phi(a^*)) \]
\[ \geq 0. \]

Therefore, with the result from b), we can get
\[ E||a_{t+1} - a^*|| \leq \frac{2}{1 + t + \ln(t)} \left( D^2 + \frac{G^2}{2\gamma^2} \right) (1 + \ln(t)). \]
Figure 2: Active Regions recovered by the proposed method, Fused LASSO and HSIC-LASSO for simulated MRI images with non-additive nonlinear responses.