Supplementary Material for “Predicting Execution Time of Computer Programs Using Sparse Polynomial Regression”

1 Proof of Theorem 3.1

Let $S$ be a subset of $\{1, 2, \ldots, p\}$ and its complement $S^c = \{1, 2, \ldots, p\} \setminus S$. Write the feature matrix $X$ as $X = [X(S), X(S^c)]$. Let response $Y = f(X(S)) + \epsilon$, where $f(\cdot)$ is any function and $\epsilon$ is additive noise. Let $n$ be the number of observations and $s$ the size of $S$. We assume that $X$ is deterministic, $p$ and $s$ are fixed, and $\epsilon|_S$ are i.i.d. and follow the Gaussian distribution with mean 0 and variance $\sigma^2$. Our results also hold for zero mean sub-Gaussian noise with parameter $\sigma^2$. More general results regarding general scaling of $n$, $p$ and $s$ can also be obtained.

Recall that the LASSO is defined as

$$\hat{\beta} = \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1. \quad (1)$$

Under the following conditions, we show that Step 1 of SPORE-LASSO, the linear LASSO, selects the relevant features even if the response $Y$ depends on predictors $X(S)$ nonlinearly:

1. The columns $(X_j, j = 1, \ldots, p)$ of $X$ are standardized: $\frac{1}{n} X_j^T X_j = 1$, for all $j$;
2. $\Lambda_{\min}(\frac{1}{n} X_S^T X_S) \geq c$ with a constant $c > 0$;
3. $\min \|X_S^T X_S\|^{-1} X_S^T f(X_S)\| > \alpha$ with a constant $\alpha > 0$;
4. $\frac{X_S^T f(X_S)}{n} < \frac{n\alpha}{2\sqrt{s+1}}$, for some $0 < \eta < 1$;
5. $\|X_S^T X_S(X_S^T X_S)^{-1}\|_{\infty} \leq 1 - \eta$;

where $\Lambda_{\min}(\cdot)$ denotes the minimum eigen value of a matrix, $\|A\|_{\infty}$ is defined as $\max_i \left[\sum_j |A_{ij}| \right]$ and the inequalities are defined element-wise.

By standard convex optimization theory, if $\bar{\beta} = (\hat{\beta}_S, \hat{\beta}_{S^c})$ with $\hat{\beta}_S \neq 0$ and $\hat{\beta}_{S^c} = 0$ satisfies

$$X_S^T (Y - X_S \hat{\beta}_S) = \lambda \text{sgn}(\hat{\beta}_S), \quad (2)$$

$$|X_{S^c}^T (Y - X_S \hat{\beta}_S)| < \lambda, \quad (3)$$

then it is the unique solution of the LASSO (1).

From Equation (2), we get

$$\hat{\beta}_S = (X_S^T X_S)^{-1} X_S^T f(X_S) + (X_S^T X_S)^{-1} [X_S^T \epsilon - \lambda \text{sgn}(\hat{\beta}_S)]. \quad (4)$$

Let $\tilde{b}$ be the the sign vector of $(X_S^T X_S)^{-1} X_S^T f(X_S)$. Set $\text{sgn}(\hat{\beta}_S) = \tilde{b}$, substitute it into equation (4), and then we have

$$\hat{\beta}_S = (X_S^T X_S)^{-1} X_S^T f(X_S) + (X_S^T X_S)^{-1} [X_S^T \epsilon - \lambda \tilde{b}]. \quad (5)$$
It can be verified that if
\[
\max \left| (X_S^T X_S)^{-1} [X_S^T \epsilon - \lambda \tilde{b}] \right| < \alpha, \tag{6}
\]
then \( \hat{\beta}_S \) defined in Equation (5) satisfies Equation (2).

Substitute \( \hat{\beta}_S \) with (5) into Inequality (3), we get
\[
\nonumber
\begin{align*}
&\left| X_S^T \left[ f(X_S) - X_S(X_S^T X_S)^{-1} X_S^T f(X_S) \right] \right| \\
&+ \left| X_S^T [I - X_S(X_S^T X_S)^{-1} X_S^T] \epsilon \right| \\
&+ \lambda X_S^T X_S (X_S^T X_S)^{-1} \tilde{b} \leq \lambda.
\end{align*}
\]
\[
\nonumber
\tag{7}
\]
By assumption,
\[
\nonumber
\left| X_S^T X_S (X_S^T X_S)^{-1} \tilde{b} \right| \leq 1 - \eta,
\]
so,
\[
\left| X_S^T \left[ f(X_S) - X_S(X_S^T X_S)^{-1} X_S^T f(X_S) \right] \right| + \left| X_S^T [I - X_S(X_S^T X_S)^{-1} X_S^T] \epsilon \right| < \lambda/2 \tag{8}
\]
is sufficient for Inequality (3).

According to the previous discussion, it suffices to prove that (6) and (8) hold with probability \( \to 1 \) as \( n \to \infty \).

We analyze (8) first. \( X_S^T [I - X_S(X_S^T X_S)^{-1} X_S^T] \epsilon \) is a Gaussian random vector with mean 0 and variance of each element at most \( n\sigma^2 \). So,
\[
\nonumber
P[\max \left| X_S^T \left[ f(X_S) - X_S(X_S^T X_S)^{-1} X_S^T f(X_S) \right] \right| > t] \leq 2(p - s) \exp \left\{ - \frac{t^2}{2n\sigma^2} \right\}.
\]
Setting \( t = \frac{\lambda n}{2} - \left| X_S^T \left[ f(X_S) - X_S(X_S^T X_S)^{-1} X_S^T f(X_S) \right] \right| \), we obtain that
\[
\nonumber
P[\text{(8) holds}] \geq 1 - 2(p - s) \exp \left\{ \frac{\left( \left| X_S^T \left[ f(X_S) - X_S(X_S^T X_S)^{-1} X_S^T f(X_S) \right] \right| - \frac{\lambda n}{2} \right)^2}{2n\sigma^2} \right\}.
\]
Set
\[
\nonumber
\lambda = \frac{2}{\eta} \left\{ \left| X_S^T \left[ f(X_S) - X_S(X_S^T X_S)^{-1} X_S^T f(X_S) \right] \right| + \kappa \sqrt{n} \log n \right\}, \tag{9}
\]
where \( \kappa \) is a constant. It is easy to see that the above probability goes to 1. From Condition 4., \( \lambda \) has the property that \( \lambda/n \leq \frac{\alpha c}{\sqrt{n} \sqrt{\log n}} \) as \( n \to \infty \).

Now we analyze (6). We have \( \left| (X_S^T X_S)^{-1} [X_S^T \epsilon - \lambda \tilde{b}] \right| \leq \left| (X_S^T X_S)^{-1} X_S^T \epsilon \right| + \left| (X_S^T X_S)^{-1} \lambda \tilde{b} \right| \).

Since \( \left\| (X_S^T X_S)^{-1} \right\|_2 \leq \frac{1}{p} \), we have the variance of each element of Gaussian vector \( \left| (X_S^T X_S)^{-1} [X_S^T \epsilon] \right| \) at most \( \frac{\alpha c}{\sqrt{n}} \).

So
\[
\nonumber
P[\max \left| (X_S^T X_S)^{-1} [X_S^T \epsilon] \right| > t] \leq 2s \exp \left\{ - \frac{nt}{\sigma^2} \right\},
\]
\[
\nonumber
\left| (X_S^T X_S)^{-1} \lambda \tilde{b} \right| \leq \frac{\sqrt{s} \lambda}{nc}.
\]
Set \( t^2 = \frac{\lambda}{\sqrt{n}} \) and set \( \lambda \) such that \( \frac{\sqrt{s} \lambda}{nc} < \alpha \) (note that the previous choice of \( \lambda \) in Equation (9) satisfies this requirement), then (6) holds with probability greater than \( 1 - 2s \exp \left\{ - \frac{\sqrt{s} \lambda}{\sigma^2} \right\} \to 1 \).

2 Full version of SPORE-FoBa algorithm
Algorithm 1 SPORE-FoBa

**Input:** data \((x_i, y_i), i = 1, \ldots, n\), the maximum degree \(d\), \(\epsilon\)

**Output:** polynomial terms \(T^{(k)}\) and the coefficients \(\beta^{(k)}\).

1. Let \(T^{(0)} = \emptyset\), \(S^{(0)} = \emptyset\)
2. let \(k = 0\) (number of terms)
3. let \(RSS^{(0)} = \sum_i y_i^2\)
4. while True do
   5. \(RSS_J = \|Y - T^{(k)} \beta^{(k)}\|_2^2\)
   6. for \(j = 1, \ldots, p\) do
      7. let \(C = \{ t : t = x_{i_1}^{d_1} \Pi_{l \in S} x_{l}^{d_l} \text{ with } d_1 > 0, d_l \geq 0, d_1 + \sum d_l \leq d \}\)
     8. // Forward step: add terms from \(C\)
      9. while True do
         10. let \(k = k + 1\)
         11. let \([t^{(k)}, \beta^{(k)}] = \arg\min_{t \in C, \beta} \|Y - [T^{(k-1)}, t] \beta\|_2^2\)
         12. let \(RSS^{(k)} = \|Y - [T^{(k-1)}, t^{(k)}] \beta^{(k)}\|_2^2\)
         13. let \(\delta^{(k)} = RSS^{(k-1)} - RSS^{(k)}\)
         14. \(T^{(k)} = T^{(k-1)} \cup t^{(k)}\)
         15. if \(\delta^{(k)} \leq \epsilon\) then
            16. \(k = k - 1\)
            17. break
      18. end if
     19. // backward step: remove terms from active set \(T^{(k)}\)
      20. while True do
         21. \(RSS_{pre} = RSS^{(k)}\)
         22. let \([t, \beta_{now}] = \arg\min_{t \in T^{(k)} \setminus \{t^{(k)}\}, \beta} \|Y - [T^{(k)}, t] \beta\|_2^2\)
         23. let \(RSS_{now} = \|Y - [T^{(k)} \setminus \{t\}] \beta_{now}\|_2^2\)
         24. \(\delta' = RSS_{now} - RSS_{pre}\)
         25. if \(\delta' > 0.5 \delta^{(k)}\) then
            26. break
      27. end if
     28. let \(k = k - 1\)
     29. let \(T^{(k)} = T^{(k+1)} \setminus \{t\}\)
     30. let \(\beta^{(k)} = \beta_{now}\)
     31. let \(RSS^{(k)} = RSS_{now}\)
      32. end while
      33. end while
   34. if Feature \(j\) is added into the active set \(T^{(k)}\) then
      35. \(S = S \cup j\)
      36. end if
   37. end for
   38. if \(RSS^{(k)} - RSS_J \leq \epsilon\) then
      39. break
   40. end if
41. end while