IDENTIFYING HETEROGENEOUS AND CONFOUNDING EFFECT VIA L_1-REGULARIZED SOFT DECISION TREE

Ran Wang

Department of Economics

University of California, Riverside

ran.wang@email.ucr.edu

October 26, 2019

ABSTRACT

In modern causal inference literature, one increasingly important problem is how to correctly and efficiently identify the common factors for resolving the confounding issues and heterogeneous treatment effect for personalized the policies when there are very high-dimensional features available. In this paper, we develop a new method called L_1-Regularized Soft Decision Tree method to identify the heterogeneous treatment effect and confounding effect under a very flexible treatment effect framework. We first convert the main problem into a nonlinear variable selection task and prove that our new method can identify the relevant subset of the relevant factors consistently. We also discuss the asymptotic theory of the soft decision tree that can be a very effective method for resolving other nonlinear variable selection issues. In our empirical experiment, we find that L_1-Regularized Soft Decision Tree can identify the relatively meaningful factors, which is better than the results returned by other variable selection methods like Lasso.

1 Introduction and Motivation

Causal inference is a powerful tool for social scientists and has been widely used in economic research, such as microeconomic theory testing, macroeconomic policy evaluation and so on. In causal inference, we often use the treatment effect, like the Average Treatment Effect (ATE) to measure the causal effect quantitatively.

Unfortunately, the method of causal treatment effect above often suffers from two main problems: the first problem is related to confounding effect, which means there exists some common factors or confounders that can lead to the distortion of the true relationship between treatment and outcomes. Considered a simple linear case related to the confounding effect:

$$\begin{aligned} Y &= \phi_0 D + \beta_y C + e, \;\; E(e|C, D) = 0, \\ D &= \beta_d C + r, \;\; E(r|C) = 0, \end{aligned} \tag{1}$$

where Y is the outcome, D is the treatment, $\beta_y C$ and $\beta_d C$ are two linear functions related to the confounder C. For example, in the classic smoking and lung cancer example, the hidden factor, like people's income or hometown, could distort the true relationship between smoking and lung cancer. To resolve this issue, researchers have proposed several ways such as the propensity score matching [Rosenbaum and Rubin [1983]] and doubly robust estimators [Lunceford and Davidian [2004]].

The second problem is the heterogeneous effect. That is, the expected outcome could depend on other factors. For example, in economic research, we usually face the problem of heterogeneity, like heterogeneous economic agents. To the same economic policy, agents with different features could have different responses. Thus, a precise estimation of the heterogeneous treatment effect is not only helpful to study the properties of agents but very important in the economic policy analysis.

Even though there are many kinds of research discussing the issues above, an important condition should be satisfied: the confounders and heterogeneous factors are available. Recently, because of

the large amount of data available, we can observe a very high-dimensional feature vector to each agent. But this does not mean two factors are available since the confounders and heterogeneous factors are often contained in one group. Thus, although we can select the meaningful variables based on the economic theories, how to select the relevant features is a very difficult problem when there are too many candidates. Additionally, it is often more reasonable to use the data-driven method than the prior knowledge to identify the important variables.

In this paper, we consider a very flexible framework to describe this issue:

$$\begin{aligned} Y_D(H, C) &= m_h(D, H) + m_c(D, C) + U, \\ D &= m_e(C) + E, \end{aligned} \tag{2}$$

where $(H, C) \in X$ are heterogeneous factors and confounders contained in one group X. To separate heterogeneous factors and confounders, we can implement a two-step variable selection as follows:

- 1 Given X, identify C by selecting the relevant variables for predicting D.
- 2 Given X, identify H and C by selecting the relevant variables for predicting Y.
- 3 Based on the two variable groups frm step 1 and 2, take the intersection to decide the C, and the remained variables are H.

Obviously, to select the correct varaibles for H and C, it is reasonible to implement a consistent nonlinear variable selection if $m_h(D, H)$, $m_c(D, C)$ and $m_e(C)$ are unknown nonlinear functions. Thus, we propose our new method called L_1-Regularized Soft Decision Tree method, which cannot only identify the heterogeneous factors H and confounders C consistently and efficiently, but can approximate the three unknown functions $m_h(D, H)$, $m_c(D, C)$ and $m_e(C)$ simutaneously.

2 Related Work

Basically, causal inference is a widely discussed topic in statistics, computer science and many social sciences like economics and politics. There are two main methods for causal inference: the Structural Causal Model (SCM), which focus on the identification of the causal structure [Pearl, Judea [2009], Pearl, Judea and Mackenzie, Dana [2018]] and treatment effect model [Imbens, Guido W. and Rubin, Donald B. [2017]], that concentrate on the estimation of the treatment's magnitude and significance. In the economic literature, the treatment effect method is usually the main way to analyze the causal effect in econometrics and applied econometrics and applied economics. In the settings of the classic random experiment, the potential outcomes framework is the backbone of many causal inference methods in the economic researches [Donald B., Rubin [1974]]. To the observational data, propensity score matching [Rosenbaum and Rubin [1983]] and doubly robust estimators [Lunceford and Davidian [2004]] are proposed and discussed for the unbiased estimation of the treatment effect given the confounders.

Recently, machine learning techniques are introduced into economics dramatically [Athey [2019],Athey and Imbens [2019]]. Because of the available large economic dataset and low-cost computational resources, machine learning methods can approximate the very complicated unknown function and handle the high-dimensional inputs issue [Varian [2014]]. For example, Random Forest, Boosting and Neural Networks methods are implemented into the economic now-casting given a variety of covariates [Richardson et al. [2018]]. Additionally, to solve the certain issues in econometrics, like GMM and instrumental variables, there are increasingly number of new methods proposed by economists based on the classic machine learning algorithms for economic inference [Lewis and Syrgkanis [2018], Hartford et al. [2016], Athey et al. [2019]].

The early machine learning research in the causal inference started in the 2000s. van der Laan and Rubin [2006] proposed the Targeted Learning framework containing the targeted MLE and super learner technique for causal effect estimation. Chipman et al. [2010] proposed the Bayesian Additive Regression Trees, which combines the decision trees in a Bayesian way and estimates the

causal effect based on the posterior distribution generated by the model. Belloni et al. [2013] and Belloni et al. [2014] discussed the estimation treatment effects given high-dimensional controls and proposed a double selection estimator for selecting relevant controls in a linear setting using Lasso. Athey and Imbens [2016] discussed an unbiased ML estimator, the Honest Tree, for testing the treatment effect. Luo and Spindler [2017] discussed a L_2-Boosting-based double selection method for the similar settings of high-dimensional controls. Wager and Athey [2018] proposed the Causal Forests for estimating the heterogeneous treatment effect with unbiasedness and asymptotic normality. Chernozhukov et al. [2017] and Chernozhukov et al. [2018] proposed a Double Machine Learning (DML) based on the Partially linear model and Residual-in-Residual method discussed by Robinson [1988], which provides a flexible estimator given a lot of attributes.

The main contribution of our paper is that we proposed a new way to identify the heterogeneous treatment effect and confounding effect. Based on the generalized partially linear framework, we combine and then extend the double selection estimator, double machine learning and causal forest method to our L_1 regularized soft decision tree, which can identify the relevant attributes for heterogeneous factors and confounders consistently given a large number of covariates under nonlinear settings.

Our paper organized as follows: In Section 3, we start with the discussion of the potential outcomes framework with related machine learning methods including double machine learning and causal forests to introduce our generalized partially linear framework and the main problem we want to solve. In Section 4, we introduced a soft decision tree and our L_1 regularized soft decision tree with its properties. In Section 5, we discuss the theories of consistent variable selection behind the L_1 regularized soft decision tree. In Section 6, a simulation study is given. Finally, an empirical study about unemployment data analysis is given based on our new method in Section 7.

3 Machine Learning in the Potential Outcomes Framework

In this section, we introduce the potential outcomes framework and the related machine learning techniques. Consider a dataset from a Randomized Control Trial (RCT) experiment with N samples.

Sample i contains a treatment indicator $D_i \in \{0, 1\}$, a response$Y_i \in \mathbb{R}$. The average treatment effect (ATE) τ is:

$$\tau = E(Y^1) - E(Y^0), \tag{3}$$

where Y^1, Y^0 are the response corresponding to treatment group and control group. The estimator $\widehat{\tau}$ is:

$$\tau = \frac{1}{\sum_{i=1}^{N} I(D_i = 1)} \sum_{i=1}^{N} Y_i \times I(D_i = 1) - \frac{1}{\sum_{i=1}^{N} I(D_i = 0)} \sum_{i=1}^{N} Y_i \times I(D_i = 0) \tag{4}$$

Practically, the RCT is often unsatisfied especially for the observational data, which is the core topic in the causal inference research.

3.1 Partially Linear Regression and Double Machine Learning

One classic case of violation to the RCT is the confoundedness. Basically, there may exist one or more common factors distorting the true dependence between the treatment D and response Y. The Partially Linear Regression (PLR) framework [Robinson [1988]] describes the issue as follows:

$$\begin{aligned} Y &= \phi_0 D + m_y(C) + e, \;\; E(e|C, D) = 0, \\ D &= m_d(C) + r, \;\; E(r|C) = 0. \end{aligned} \tag{5}$$

In this framework, Y is outcome, D is a regressor with the structural parameter ϕ_0, C represent confounders. Also, $m_y(C)$, $m_d(C)$ are nonlinear functions.

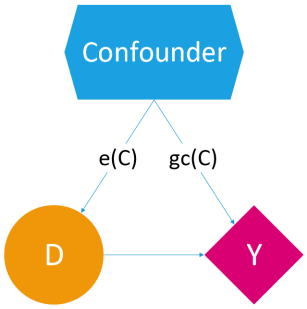

Figure 1: Neyman Partially Linear Model

Figure 1 is a simple example of PLR. Suppose we study the treatment effect of smoking on lung cancer. Because there exist many hidden factors affecting smoking and lung cancer simultaneously, like income, we cannot check the causality via regress treatment variable, smoking, on the outcome of cancer. That is the problem of endogeneity induced by common hidden factors.

In economic research, many studies suffer from this problem, which boosts the development of instrument variables (IV) theory. In IV theory, an economist could choose some valid and effective instrument variables to substitute regressors so that the estimators of structure parameters are consistent with the true value.

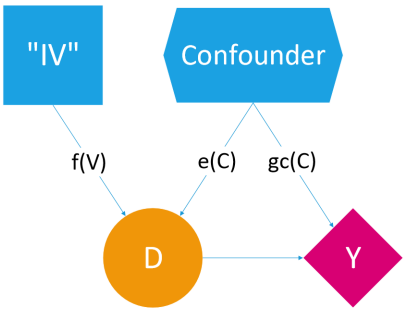

Figure 2: Instrument Variable Method

Precisely, if we select V as a IV for D, then we have:

$$\begin{aligned} Y &= \widehat{D}\phi_0 + m_y(C) + e \\ \widehat{D} &= \widehat{f}(V) \end{aligned} \tag{6}$$

According to the theory of IV, since V is independent with C but correlated with D, the estimated value $\hat{\phi}_0$ is consistent to true ϕ_0.

Hypothesitically, if confounders C are observable and accessible, we can directly remove their effects. In Robinson [1988], the author proposed a double residual semiparametric regression estimator based on the PLR framework. Recently, Chernozhukov et al. [2017] proposed a double machine learning (DML) framework. The two methods share very similar intuition that they approximate the nonlinear functions $m_y(C)$ and $m_d(C)$ via nonparameteric regression or machine learning methods. For example, the DML estimator works as follows:

$$\begin{aligned} Y &= \hat{r}\phi_0 + \hat{m}_y(C) + e \\ \hat{r} &= D - \hat{m}_d(C) \end{aligned} \tag{7}$$

In Chernozhukov et al. [2017], this is also called orthogonal machine learning. That is, to estimate the unknown functions m_y, m_d, we construct score function that satisfies Neyman orthogonal condition and identification condition:

$$\begin{aligned} \partial_\eta E\left(\psi(D, \phi_0, \eta)\right)|_{\eta=\eta_0} &= 0 \\ E\left(\psi(D, \phi_0, \eta_0)\right) &= 0 \end{aligned} \tag{8}$$

where ϕ_0 is structure parameter and $\eta_0 = (m_y, m_d)$ is the nuisance "parameter".

Additionally, the DML paper proposed to separate the sample into k subgroups. Then choose one subgroup to estimate nuisance parameters η_0 and use the other $k-1$ groups to estimate structural parameter θ_0 so that we can control the VC dimension to prevent over-fitting of the machine learning algorithm and estimate the unbiased structural parameter.

Finally, to prevent from loss of efficiency of separating samples, we can switch subgroups k times to do the estimation and then take the average value among all the estimated values, which is called cross-fitting. The DML paper also proved that when a machine learning algorithm can approximate functions at convergence rate of $n^{-1/4}$ or better, the structure parameter ϕ_0 is consistent and asymptotic normal by the estimating process above.

Similarly, this framework can be applied into the estimation of binary treatment effect:

$$\begin{aligned} Y &= m_y(D, C) + r \;\; E(e|D, C) = 0 \\ D &= m_d(C) + r, \;\; E(r|C) = 0 \end{aligned} \tag{9}$$

Based on this framework, the treatment effect is:

$$\phi_0 = E(m_y(1, C) - m_y(0, C)).$$

If we use DML to estimate average treatment effect, we can construct a score function that satisfies the two conditions discussed above:

$$\psi(D, \phi, \eta) = m_y(1, C) - m_y(0, C) + \frac{D(Y - m_y(1, C))}{m_d(C)} - \frac{(1 - D)(Y - m_y(0, C))}{1 - m_d(C)} - \phi \tag{10}$$

Furthermore, Mackey et al. [2018] extended the Neyman orthogonal condition into kth order orthogonal condition such that estimator structural parameter could also be consistent even though the machine learning estimator has a slower convergence rate at $n^{-1/(2p+2)}$.

3.2 Heterogeneous Treatment Effect and Causal Forest

The heterogeneous treatment effect has been widely studied. But one drawback is that not only these methods cannot fully catch the heterogeneity, but there is no asymptotic theory for the hypothesis test. The Causal Forest Wager and Athey [2018] is the first machine learning method that can catch the heterogeneous treatment effect and provides the asymptotic theories for inference.

Causal Forest is based on heterogeneous treatment effect framework:

$$\begin{aligned} Y &= m_y(D, X) + U \\ P(D = 1) &= e_0(X) \end{aligned} \tag{11}$$

where Y is outcome and D is a binary treatment. X is a group of heterogeneous factors. $e(X)$ is propensity score and $m_y(D, X)$ is a nonlinear function leading to heterogeneous treatment effect. Thus, the treatment effect is $\tau(x) = E[Y_i^1 - Y_i^0 | X_i = x]$ and an unbiased estimator is based on the Inverse Propensity Score Weighting (IPSW):

$$Y_i^* = Y_i \left(\frac{D_i}{e(X_i)} - \frac{1 - D_i}{1 - e(X_i)} \right) = \begin{cases} \frac{Y_i^1}{e(X_i)} & if\ D_i = 1 \\ -\frac{Y_i^0}{1-e(X_i)} & if\ D_i = 0 \end{cases} \tag{12}$$

Given the new samples (Y_i^*, X_i), the heterogeneous treatment effect $\tau(x)$ can be estimated by regressing Y^* on X via linear or nonlinear regression. Athey and Imbens [2016] is the first paper to explore the application of tree method in the treatment effect estimation and introduce the honest tree method which could estimate heterogeneous treatment effect unbiasedly. The key to the honest tree is sample splitting. That is, one group of samples is selected to grow a tree and the other group

is used for calculating the predicted value in each leaf. The authors considered using a tree-based method, the causal forest, to identify this effect. They have proved the causal forest can identify the unbiased heterogeneous treatment effect and also apply the hypothesis test on the effect. Wager and Athey [2018] discussed the Causal Forests such that the heterogeneous treatment effect is estimated consistently with asymptotic normality.

3.3 Identify Heterogeneous Effect and Confounding Effect in Treatment Effect

Based on the discussion in section 3.1 and 3.2, we can see that given the feature vector X or confounder C, we can estimate the unbiased treatment effect and identify the heterogeneous treatment effect. But two methods do not only need an indispensable assumption that the information of the feature vector X or confounder C should be available.

Suppose we have a very high-dimensional vector containing all the available features, it is not reasonable to treat all the features as the heterogeneous factors X or C. In other words, we should know that what are the features relevant to the heterogeneous effect, what are the variables in the group of confounders and what are the variables irrelevant.

In DML or the Causal Forests, they treat all the factors in one group. Precisely, the DML method estimates the unbiased treatment effect via deducting all the confounding effects. The Causal Forest considers all features to estimate the heterogeneity in the data. Thus, these two methods probably underestimate and overestimate the purely heterogeneous treatment effect. When we treat all the factors as one group, we not only estimate heterogeneous treatment effects inaccurately but could misunderstand the heterogeneity among data.

To solve it, we consider a Generalized Partially Linear Regression Framework (GPLR) illustrated in Figure 3.

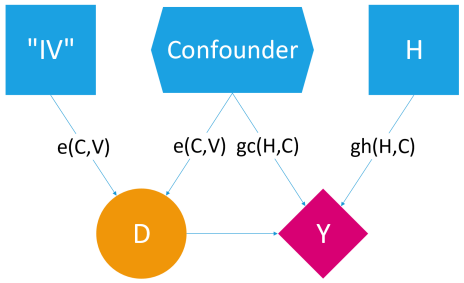

Figure 3: Generalized Partially Linear Regression

The related equations are:

$$\begin{aligned} Y_D(H, C) &= m_h(D, H) + m_c(D, C) + U, \\ P(D = 1) &= e(C, V), \end{aligned} \tag{13}$$

where H, V and C represent heterogeneous factors, instrument variables and confounders. Also, $h(D, H)$, $c(D, C)$ and $e(C, V)$ are nonlinear functions.

Thus, the core problem is to select H, V and C from one group of variables X. Let us go back to smoking and lung cancer example. At first, income could be one of the confounders affecting smoking and lung cancer simultaneously. Also, the body's health should belong to the group of heterogeneous factors that are only related to lung cancer. At last, the price of the cigarette could be an instrument variable to smoking.

A related problem was discussed in Belloni et al. [2013]. The paper suggested a double selection method for identifying the confounders C under a linear setting:

$$\begin{aligned} Y_D &= \phi_0 D + \beta_y C + e, \\ D &= \beta_d C + r. \end{aligned} \tag{14}$$

To select the correct confounders C given a high-dimensional vector X, two Lasso regressions are implemented on the equations of Y_D and D. Then, the selected variables are combined to decide the correct C. We can see that equations (12) are a simple linear version of the GPLR framework. In GPLR, we not only need to identify three factors (H, V and C), but the functional form are unknown ($h(D, H)$, $c(D, C)$ and $e(C, V)$). Thus, based on our Generalized Partially Linear Model, we suggest the following process for identifying factors:

- 1. Estimate propensity score $e(C, V)$ and identify the factors including confounders C and IVs V. The remaining subset is H (and irrelevant factors R).
- 2. Regress $Y_D(H, C)$ on all the factors to select the confounders C and heterogeneous factors H.
- 3. Regress $Y_D(H, C)$ on C to estimate $\tilde{m}_c(D, C)$
- 4. Estimate unbiased heterogeneous ATE $\tau(H)$ based on score function $\psi(D, \theta, \eta) = \tilde{m}_c(1, C) - \tilde{m}_c(0, C) + \frac{D(Y-\tilde{m}_c(1,C))}{e(C,V)} - \frac{(1-D)(Y-\tilde{m}_c(0,C))}{1-e(C,V)} - \tau(H)$

Now the problem is how to identify H, C, V given the unknown functions. To solve that, first, we need a nonlinear method that can approximate the unknown functions. Second, we need a consistent variable selection method based on the nonlinear approximator. In other words, we need consistent nonlinear variable selection methods.

In this paper, we introduce the L_1 regularized soft decision tree method to resolve these issues, which cannot only fit the unknown functions but select the correct variables simultaneously. We will discuss our methods in detail from the next chapter.

4 Nonlinear Regression via Soft Decision Tree

From this chapter, we start to introduce our main contribution. Firstly, we introduce the Soft Decision Tree, a very flexible nonlinear regression method. Second, we discuss a L_1 regularized

nonlinear regression method which can consistently select variables in the nonlinear setting. Then we combine two methods to construct the L_1 regularized soft decision tree method. We also give the related proofs of the oracle properties supporting the consistent variable selection.

4.1 From Classic Hard Decision Tree to Soft Decision Tree

In section 4.1 we discuss the Soft Decision Tree (SDT) method. Decision Tree (DT) is a classic learning algorithm in statistics and computer science. The Classification And Regression Tree (CART), a revised decision tree method, has been widely used in data mining, machine learning, and other related research fields. We mainly focus on the CART in our paper.

A decision tree is constructed by one root node, many internal nodes, and leaf nodes. Figure 4 illustrates the typical structure of a node.

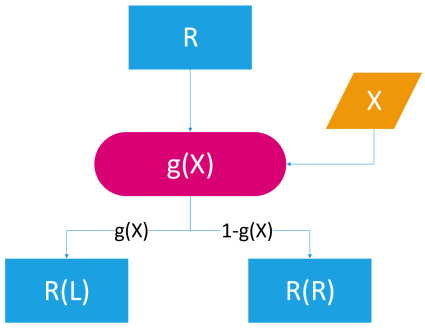

Figure 4: Root Structure in Decision Tree

And the formula of the node is:

$$R_b(x) = R_b^L(x)g_b(x) + R_b^R(x)(1 - g_b(x)), \tag{15}$$

where $g_b(x)$ is a threshold function for bth node. In the classic decision tree, $g_b(x)$ is usually an indicator function such that $g_b(x) = 0 or 1$. For this case we could have binary output:

$$R_b(x) = \begin{cases} R_b^L(x) & if\ g_b(x) = 1 \\ R_b^R(x) & if\ g_b(x) = 0 \end{cases} \tag{16}$$

In terms of the $g_b(x)$, the decision tree is "hard". Conversely, a decision tree could also be "soft" when $g_b(x)$ has a continuous output like $0 < g_b(x) < 1$. Typically, we often choose the logistic function $g_b(x) = \frac{1}{1+\exp(-(w_b^T x + w_{b_0}))}$, where w is the weight vector. For this case we have a average output between R_b^L and R_b^R:

$$\begin{aligned} R_b(x) &= R_b^L(x) g_b(x) + R_b^R(x)(1 - g_b(x)) \\ g_b(x) &= \frac{1}{1 + \exp(-(w_b^T x + w_{b_0}))} \end{aligned} \tag{17}$$

Although a hard decision tree grows by following the recursive splitting rule, there exists more than one way to construct a soft decision tree. Jordan and Jacob [1994] discussed the soft decision tree as a Hierarchical Mixtures of Experts (HME) method for the first time. To HME, we can designate the structure of the tree and then optimize the parameters via the Expectation-Maximization (EM) method.

Figure 5 shows a two-level Hierarchical Mixtures of Experts. First, in the bottom nodes, there are four functions, called the expert network, which is equivalent to the leaf node in the decision tree. They could be constant or some function of input X. Practically, a linear function is enough to catch the nonlinearity among data. To regression problem, these experts will give real value output $f(x)$. To binary or multiple classification problems, the output could be transformed into probability via logistic or softmax functions. Second, we can see three threshold functions $g(X)$, $g_1(X)$ and $g_2(X)$. These functions take input X in and calculate the probability that we need to choose the left node at

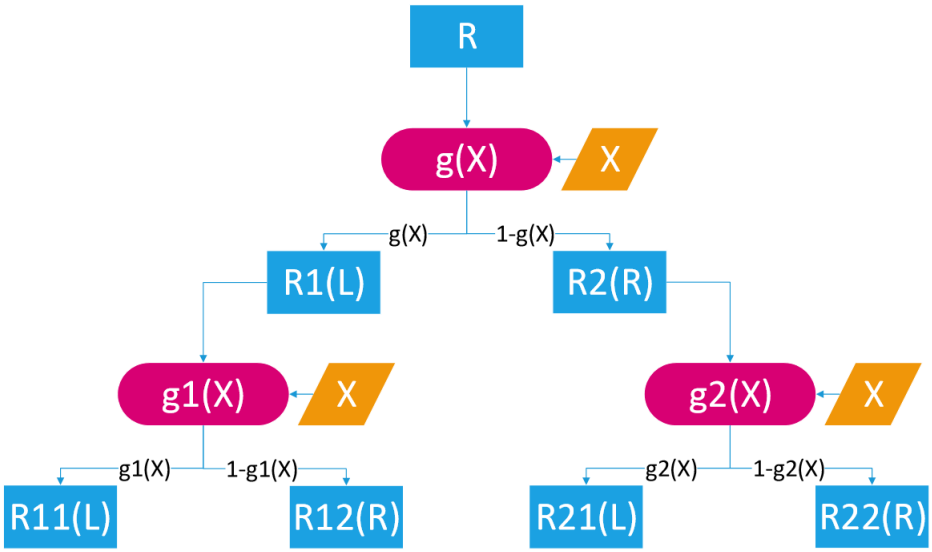

Figure 5: A Graph of Hierarchical Mixtures of Experts with Two Levels

the next level. They are also called gating functions. As we mentioned earlier, the common choice of gating function is the logistic regression function $g_b(x) = \frac{1}{1+\exp(-(w_b^T x + w_{b_0}))}$.

To calculate the output for root R in level one, we need to calculate all the probabilities for gating function $g(X)$, $g_1(X)$ and $g_2(X)$. Then, we use $R_b(x) = R_b^L(x)g_b(x) + R_b^R(x)(1 - g_b(x))$ to calculate the outputs for all the outputs of roots in high level. Finally, we accumulate all the outputs recursively to give the output of R.

The HMC covers all the intuitions behind the soft decision tree. First, this method considers regression as a binary problem recursively. To any given space of variables X, this system is bisecting this space into two subspaces. Second, it introduces a soft gating function, like the soft decision tree, instead of a hard step function, like the hard decision tree. At last, different as hard decision tree which grows a whole tree according to some information criteria, soft decision tree first designate the structure and parameters of a tree. Then, optimize all the parameters based on

maximizing the likelihood function. Since all the parameters are adjusted simultaneously, a soft decision tree often outperforms a hard decision tree in many cases. To generalized we could add more nodes into a soft decision tree to improve the approximating power and decrease the bias.

To optimize the parameters in the soft decision tree, the likelihood function needs to be constructed. According to the definition of soft decision tree, the probability of y given x and all the parameters w is:

$$P(y|x,w) = \sum_{s=1}^{S} \prod_{p \to s} g_p(x; w_p) P_s(y|x) \tag{18}$$

Thus, the likelihood function of soft decision tree is:

$$L(y,x;w) = \sum_{i=1}^{N} Log\left(P(y_i|x_i,w)\right) = \sum_{i=1}^{N} Log \sum_{s=1}^{S} \prod_{p \to s} g_p(x_i; w_p) P_s(y_i|x_i) \tag{19}$$

Assume that y follows a certain distribution P in every node, like Gaussian distribution, The final likelihood function is:

$$L(y,x;w,\mu,\Sigma) = \sum_{i=1}^{N} Log \sum_{s=1}^{S} \prod_{p \to s} g_p(x_i; w_p) P_s(y_i; \mu_s, \Sigma_s), \tag{20}$$

where μ_s, Σ_s are the mean and covariance matrix for node s.

Unforetrunately, the optimal solution based on the likelihood function does not have a closed form and the numerial methods are often considered. According to Jordan and Jacob [1994], we can maximize the likelihood function via Expectation-Maximization (EM). In the EM method, we consider a complete likelihood function $L_c(y,x;w,\mu,\Sigma)$ with the implicit variables z_s to all the nodes:

$$L_c(y,x;w,\mu,\Sigma) = \sum_{i=1}^{N} \sum_{s=1}^{S} z_s Log \prod_{p \to s} g_p(x_i; w_p) P_{node}(y_i; \mu_{node}, \Sigma_{node}). \tag{21}$$

The Expectation is of the complete likelihood function is:

$$\begin{aligned} Q(y, x; w, \mu, \Sigma) &= E_z(L_c(y, x; w, \mu, \Sigma)) \\ &= \sum_{i=1}^{N} \sum_{s=1}^{S} E(z_s) Log \prod_{p \to s} \sigma_p(x_i; w_p) P_s(y_i; \mu_s, \Sigma_s) \end{aligned} \tag{22}$$

Intuitively, instead of optimizing the original likelihood function, the EM method optimizes the lower bound of the likelihood function, which leads to a more efficient way to update the parameters especially when $P_s(y_i; \mu_s, \Sigma_s)$ belongs to the exponential family. Considered more general cases, we can optimize the original likelihood function via gradient descend or Newton method. In our paper, we choose Gradient Descend (GD) to optimize the parameters in the soft decision tree. We show that we can maximize the likelihood function by minimizing the loss function of the soft decision tree. Precisely, we have the following result 1.

Result 1: assume $P_s(y_i; \mu_s, \Sigma$ has the same covariance matrix in each node, optimize the loss function $\sum_{i=1}^{N}(y_i - \sum_{node=1}^{S^2} \alpha_{node}\mu_{node})^2$ is equivalent to optimize the lower bound of the original likelihood function $L(y, x; w) = \sum_{i=1}^{N} Log \sum_{s=1}^{S} \prod_{p \to s} g_p(x_i; w_p) P_s(y_i|x_i)$.

Based on this result, we not only optimize a penalized loss function very easy but supports to the discussion about the asymptotic properties in the following sections.

4.2 Soft Decision Tree As a Nonparametric Kernel Regression

In this section, we discuss the property of the soft decision tree. We will show the soft decision tree method is equivalent to a nonparameteric kernel regression under conditions. Without loss of generalization, we set $f(x) = \bar{y} = \mu$ for each leaf node. The formula of a soft decision tree is:

$$\begin{aligned}
\mu(x) &= \sum_{s=1}^{S} \left(\prod_{p \to s} g(x; w_p) \right) \times \mu_s \\
&= \sum_{s=1}^{S} \left(\prod_{p \to s} \frac{1}{1 + exp(-(w_p x + w_{p0}))} \right) \times \mu_s \\
&= \sum_{s=1}^{S} \left(\prod_{p \to s} \frac{1}{1 + exp(-(w_p x + w_{p0}))} \right) \times \mu_s
\end{aligned} \tag{23}$$

where S is the number of leaf node in a tree, and w_p, w_{p0}, μ_s are unknown parameters. $p \to s$ means that all the nodes locating in the path to this final node are considered.

Now rewrite the $g(x; w_p, w_{p0})$ as:

$$\begin{aligned}
g(x; w_p, w_{p0}) &= \frac{1}{1 + exp(-(w_p x + w_{p0}))} \\
&= \frac{exp(1/2(w_p x + w_{p0}))}{exp(-1/2(w_p x + w_{p0})) + exp(1/2(w_p x + w_{p0}))}
\end{aligned} \tag{24}$$

Next, let us consider another possible way to derive this function. Suppose we have a Gaussian kernel function $exp(-\beta(x - c_0)^2)$. According to the Gaussian Mixture Model, we can have:

$$g(x; \beta, c) = \frac{exp(-\beta(x - c_0)^2)}{exp(-\beta(x - c_0)^2) + exp(-\beta(x - c_1)^2)}, \tag{25}$$

where β, c_0, c_1 are parameters for the Gaussian kernel function.

Arrange and rewrite this function, we have:

$$\begin{aligned} g(x;\beta,c) &= \frac{exp(-\beta(x-c_0)^2)}{exp(-\beta(x-c_0)^2)+exp(-\beta(x-c_1)^2)} \\ &= \frac{exp(-\beta(x^2-2c_0x+c_0^2))}{exp(-\beta(x^2-2c_0x+c_0^2))+exp(-\beta(x^2-2c_1x+c_1^2))} \\ &= \frac{exp(2\beta(c_1/2+c_0/2)x+\beta c_1^2/2-\beta c_0^2/2)}{exp(2\beta(c_1/2+c_0/2)x+\beta c_1^2/2-\beta c_0^2/2)+exp(-(2\beta(c_1/2+c_0/2)x+\beta c_1^2/2-\beta c_0^2/2))} \end{aligned} \tag{26}$$

Next, let $2\beta(c_1/2+c_0/2)=w_p$ and $\beta c_1^2/2-\beta c_0^2/2=w_{p0}$, we have:

$$\begin{aligned} g(x;\beta,c) &= \frac{exp(2\beta(c_1/2+c_0/2)x+\beta c_1^2/2-\beta c_0^2/2)}{exp(2\beta(c_1/2+c_0/2)x+\beta c_1^2/2-\beta c_0^2/2)+exp(-(2\beta(c_1/2+c_0/2)x+\beta c_1^2/2-\beta c_0^2/2))} \\ &\to \frac{exp(-1/2(w_px+w_{p0}))}{exp(-1/2(w_px+w_{p0}))+exp(1/2(w_px+w_{p0}))} \\ &= g(x;w_p,w_{p0}) \end{aligned} \tag{27}$$

Thus, a logistic function can be represented by two kernel functions with same window width β and the symmetric centers c_0, c_1. We can rewrite the soft decision as a kernel regression function with finite number of kernels:

$$\begin{aligned} \mu(x) &= \sum_{s=1}^{S}\left(\prod_{p\to s} g(x;w_p)\right)\times\mu_s \\ &= \frac{\sum_{s=1}^{S}K(x;\beta_s,c_s)\mu_s}{\sum_{s=1}^{S}K(x;\beta_s,c_s)}. \end{aligned} \tag{28}$$

Based on this conclusion, we can analyze the unbiasedness of the soft decision tree based on mixture models with finite kernels. There we give the result 2 to show the unbiasedness of the soft decision tree:

Result 2: given the following assumptions:

- 1 the unknown function $f(x)$ is Lipschitz continuous:

$$|f_0(x) - f_0(x')| \leq D|x - x'|, \tag{29}$$

  where D is the Lipschitz constant.

- 2 $Var(c_s) \neq 0$.

- 3 $|x|$ in finite.

The the bias of the soft decision tree $\mu(x) = \frac{\sum_{s=1}^{S} K(x;\beta_s,c_s)\mu_s}{\sum_{s=1}^{S} K(x;\beta_s,c_s)}$ is bounded by:

$$Bias = E|f_0(x) - f(x)| \leq \frac{1}{\sqrt{2}} D S^{-C\frac{\beta}{p}(1/2-d_m^2)} \to 0, \tag{30}$$

where $C = 1/log(2)$, p is the dimension of the input x.

Figure 6 illustrates the connection between soft decision tree and kernel regression. Since each split can generate two kernels and the total number of kernels depends on the number of leaf nodes. Also, the center of every kernel is estimated. In the kernel regression, the sample size decides the number of kernels and each kernel is centered on a certain sample. Thus, two methods share similar asymptotic properties. But the soft decision tree can be more adaptive to the sample than the kernel regression since it decides the position of the kernel on the distribution of samples.

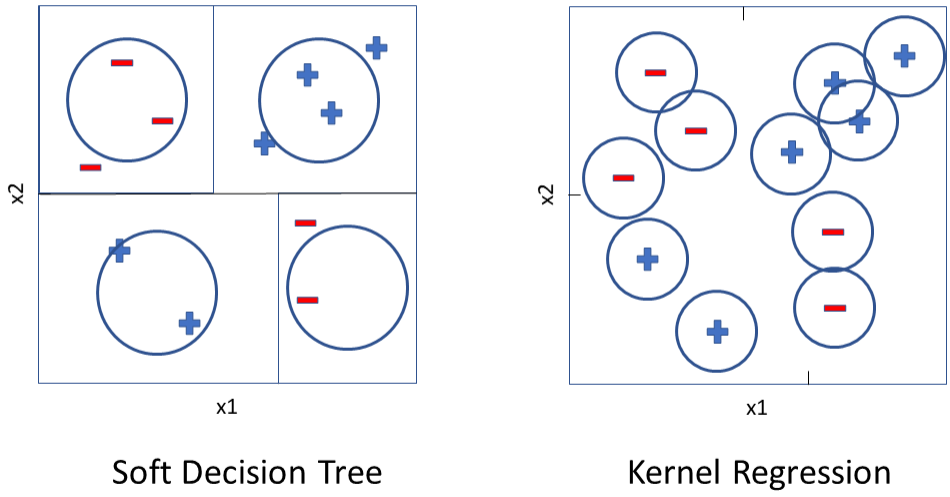

Figure 6: The Connection between Soft Decision Tree and Kernel Regression

4.3 Asymptotic Properties of Soft Decision Tree

After constructing the bound of the bias in section 4.2, we discuss the asymptotic properties of the soft decision tree in this section. In section 4.1, we discussed that maximizing the likelihood function is equivalent to minimizing the loss function in the soft decision tree. Based on the result in section 4.2, the soft decision tree can approximate any Lipschitz continuous function. Thus, we can conclude that the likelihood function of the soft decision tree can approximate the true likelihood function based on the true data generating process and the results based on the soft decision tree's likelihood function can be generalized to the true likelihood function.

The remaining question is: could the MLE estimator $\hat{f}$ be consistent to the true parameters f for the soft decision tree? To solve it, we need to go back to the formation of the soft decision tree model:

$$\begin{aligned} E(y|x) &= \frac{\sum_{s=1}^{S} K(x;\beta_s,c_s)\mu_s}{\sum_{s=1}^{S} K(x;\beta_s,c_s)} \\ &= \sum_{s=1}^{S} \alpha_s(x;\theta)\mu_s(x) \end{aligned} \tag{31}$$

And the probability density function is:

$$f(y|x) = \sum_{s=1}^{S} \alpha_s(x;\theta) P_s(y|x;\theta) \tag{32}$$

The likelihood function is:

$$L(y,x;\theta) = \sum_{i=1}^{N} Log \sum_{node=1}^{2^S} \alpha_{node}(x_i;\theta) P_{node}(y_i|x_i;\theta) \tag{33}$$

First, we consider the consistency of the MLE estimator of parameters. According to the proof by Kiefer and Wolfowitz [1956], the consistence of the nonparametric MLE needs four conditions:

Define a distance $D(.,.)$ on G:

$$D_{KW}(G_1,G_2) = \int_{\Theta} |G_1(\theta) - G_2(\theta)| exp(-|\theta|) d\theta \tag{34}$$

- 1 Identifiability: to different parameters G and G^*, if $F(x;G) = F(x;G^*)$, the $D(G,G^*) = 0$.

- 2 Continuity: $lim_{G \to G_0} f(x;G) = f(x;G_0)$,

- 3 Finite K-L Information: $E(log(f(X; B_\epsilon(G))/f(X; G^*)))^+ \leq \infty$ where $G \neq G^*$ and $B_\epsilon(G)$ is an open ball around G with radius of ϵ.

- Compactness: on space G

Based on the satisfaction of the assumptions above, the nonparametric MLE is strongly consistent, or $D_{KW}(\hat{G}, G^*) \to 0$

Unfortunately, the four conditions cannot be verified easily, especially for finite KL information. Pfanzagl [1988] suggested another way to prove the consistency of MLE with only the satisfaction of conditions 1 and 2. Specifically, Pfanzagl introduced a new inequality:

$$E^* log(1 + u(f^*/f - 1)) \geq 0$$

where $u \in (0, 1)$ and equality holds iff $f^* = f$. Based on this, assume that the mixture model is identifiable (condition 1) and condition 2 is satisfied, we have:

$$D_{KW}(\hat{G}_n, G^*) \to 0$$

Now we need to check if the soft decision tree model satisfies the two conditions or not. First, as a generalized mixture model, the soft decision tree also suffers from the label switching problem, which means giving an object density, there exists more than one parameter giving the same density. But since the object is density, not the parameters, we can revise the condition 1 to local identifiable, which means at an open set of parameters around one local optimal parameter, condition 1 is satisfied. The second problem is the continuity, which means the density function is not continuous on all the parameters, like the variance σ^2 in Gaussian density. In Chen [2017], there is a lemma that there exists a finite number σ^2 such that $(\hat{G}, \sigma^2)$ is a global minimum. Practically, we can try different initially value such that the σ^2 can converge to a finite number.

To sum up, we have the result 3.

Result 3: Suppose the data generate from a mixture density $f(x, y; \theta)$. Also, the above condition 1 to 4 of the likelihood function, score function and information matrix are satisfied, we have:

$$\sqrt{n}(\hat{\theta} - \theta) \sim N(0, \Sigma_\theta),$$

where $\Sigma_\theta = I^{-1}(\theta)$ is the asymptotic covariance matrix based on MLE estimator.

And according to delta method, we also have:

$$\sqrt{n}(\hat{\mu}(x) - \mu(x)) \sim N(0, J_\theta(x)^T \Sigma J_\theta(x)).$$

Back to the likelihood function of the Soft Decision Tree, since we have showed that the Soft Decision Tree can approximate any Lipchitz continuous function consistently, the likelihood function of the Soft Decision Tree can converge to the trun likelihood function. Thus, the asymptotic resuls in Result 3 will hold.

5 Variable Selection via Soft Decision Tree

5.1 Oracle Properties for Nonlinear Variable Selection

Variable selection is a significant topic in statistics, economics and other modeling problems that suffer from the ultra-high dimensional input space. In linear variable selection case, L_1 regularization methods, like LASSO, Elastic Net and L_2-Boosting, can select important variables simultaneously. But there are few studies about the variable selection in the nonlinear case, like tree model, kernel regression and neural network. Some researchers (L. Rosasco et al, 2010) considered a regularization approach to nonlinear variable selection via penalizing the L_2 norm of the first-order derivative on kernel machine. This idea shed a light on variable selection on any nonlinear regression methods.

We start from the linear regression model. Suppose the sample set $(y_i, x_i), i = 1, ..., N$ are generated from a the following DGP:

$$y_i = \beta x_i + u_i, i = 1, 2, ..., N$$

where β is a p dimensional vector and $u_i, i = 1, ..., N$ are series of random noise following a normal distribution $N(0, \sigma^2)$. Obviously, if we consider the loss function $L = \sum_{i=1}^{N}(y_i - \beta x_i)^2$, the OLS estimator of β is:

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y$$

This is the optimal estimator when all the settings above are satisfied. Additionally, if we add another condition:

$$\beta_p = \begin{cases} 0 & if\ p \geq r \\ constant & if\ p < r \end{cases}$$

That is, β is now a sparse vector. In this case, the OLS estimator is not the optimal one since it usually every element a non-zero value. To solve it, we need to select the best subset of coefficients β_p that are related to y and give other coefficients zero values. This is related to variable selection. The most straight forward method is called the best subset, which considers all the combinations of subsets of variables in the model. Unfortunately, this is a time-consuming process especially for the high dimensional β. To solve this, the idea of the Lasso method has been introduced in Tibshirani [1996], which suggests penalizing the L_1 norm of coefficient vector β. Thus, the new loss function is:

$$L_{LASSO} = \sum_{i=1}^{N}(y_i - \beta x_i)^2 + \lambda ||\beta||_1$$

If we assume each input variable is independent with other variables, the solution of LASSO have the following closed form:

$$\hat{\beta}_{LASSO} = \widehat{\beta}_{OLS}\left(1 - \frac{\lambda}{||\widehat{\beta}_{OLS}||_2}\right)^+$$

The last 20 years witnessed the dramatic increase in studies about LASSO based variable selection. Unfortunately, the Lasso type variable selection method may not always give the correct subgroup of variables unless some important conditions and assumptions are satisfied. Thus, could L_1 penalty term select correct variables consistently have been an indispensable question in the last 10 years. To solve this problem, many studies explored the oracle properties of variable selection methods [Fan and Li [2001],Zou [2006]]:

Theorem (Oracle properties): A variable selection estimator satisfies oracle properties if and only if two following conditions are satisfied:

- 1 Consistency in variable selection: $\lim_n P(A_n^* = A) = 1$.
- 2 Asymptotic normality: $\sqrt{n}(\hat{\beta}_A^{*(n)} - \beta_A^*) \rightarrow_d \ N(0, \times\Sigma^{-1})$, where A is a set including all the relevant variables and Σ is the covariance matrix to $\hat{\beta}_A^{*(n)}$.

As a consequence, so many methods with oracle properties were developed, such Adaptive LassoZou [2006], SCADFan and Li [2001] and so on. The main intuition about these methods are similar. That is, to different coefficients, the penalty terms could give different weights. For example, the adaptive LASSO's loss function is:

$$L_{ADALASSO} = \sum_{i=1}^{N}(y_i - \beta x_i)^2 + \lambda\sum_{p=1}^{P}\frac{1}{w_p}|\beta_p|$$

where $w_p = |\hat{\beta}_{OLS}|^\gamma$.

Adaptive LASSO gives a larger penalty to the coefficient with smaller OLS estimated value and smaller penalty to the coefficient with a larger value. Zou [2006] had proved that under some conditions on γ and λ, adaptive LASSO could select correct variables consistently. Thus, these "Oracle" methods provide the insurance to variable selection theoretically.

But to the nonlinear model, variable selection, to our knowledge, has not been widely discussed. There are only a few researches about this problem. In the next section, we will explore the oracle properties of L_1 regularized nonlinear regression.

Now we move into the nonlinear model. Consider a data generate process:

$$y = f(x) + u$$

Suppose we have a set of feature variable X, where $x \subseteq X$. If we regress y on all the X via some nonlinear regression methods, we have:

$$y = g(X) + v$$

Since we have $x \subseteq X$, some variables in X should be irrelevant with y. Suppose the DGP is $y = f(x_1) + u$. Now x_1 is the set of relevant variables and x_2 is the set of irrelevant variables. Thus, we have new nonlinear regression model:

$$y = g(x_1, x_2) + v$$

Since x_2 is irrelevant with y, an obvisous consequence is the partial derivative $\frac{\partial g(x_1,x_2)}{\partial x_2}$ should be zero. And then $\frac{\partial g(x_1,x_2)}{\partial x_1} = \frac{\partial f(x_1)}{\partial x_1}$ to any value of x_2. Based on this, we can select variable via

penalizing the norm of first derivative of $\frac{\partial g}{\partial X}$ when the functions given by nonlinear regression methods are differentiable:

$$Min:\ Loss = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda \sum_{p=1}^{P} \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial f}{\partial x_p}|_{x_p=x_{p,i}}\right)^2} \tag{35}$$

where the penalty term $\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial f}{\partial x_p}|_{x_p=x_{p,i}}\right)^2}$ is the discrete version of the L_2 norm of first derivative function:

$$\left|\left|\frac{\partial f}{\partial x_p}\right|\right|_2 = \sqrt{\int_{x_p}\left(\frac{\partial f}{\partial x_p}\right)^2 dx_p}$$

In linear model selection, we often select the best subgroup of variables and then give a consistent estimation based on selected variables, which is called post-selection analysis. In nonlinear variable selection, we could also follow this process. Thus, at first, we need to prove that the nonlinear variable selection method with L_1 penalty has oracle property and then the nonlinear estimator can also be an asymptotic normal estimator in the post-selection analysis.

Back to the loss function of nonlinear regression with L_1 penalty term:

$$Min:\ Loss = \frac{1}{N}\sum_{i=1}^{N}(y_i - \tilde{f}(x_i))^2 + \lambda \sum_{p=1}^{P} \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial \tilde{f}}{\partial x_p}|_{x_p=x_{p,i}}\right)^2} \tag{36}$$

According to the first order condition we have:

$$\frac{2}{N}\sum_{i=1}^{N}(y_i - \tilde{f}(x_i))\frac{\partial \tilde{f}(x_i)}{\partial \theta_s} = \lambda \sum_{p=1}^{P} \frac{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial \tilde{f}}{\partial x_{p,i}}\right)\frac{\partial \tilde{f}}{\partial x_{p,i}\theta_s}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial \tilde{f}}{\partial x_{p,i}}\right)^2}} \tag{37}$$

where $\tilde{f}(x)$ is the L_1 regularized nonlinear regression estimator.

Many variable selection methods feature the oracle properties via the adaptively weighted penalty. For example, the penalty term of adaptive Lasso is:

$$\sum_p \hat{w}_p|\beta_p|$$

where $\hat{w}_p = \frac{1}{|\hat{\beta}_p|^\gamma}$. The intuition is that the larger the estimated value is, the smaller the weight of the penalty is. As a consequence, all the β with small values will be shrunk to zero, when λ and γ satisfy some conditions [Zou [2006]]. SCAD also work in a very similar way as an adaptive Lasso. In our paper, we introduce the idea of adaptive lasso into the nonlinear variable selection.

Proposition 1: suppose the data generating process is $y = g(x) + \epsilon$, where $g(x)$ is a first order lipschitz continuous function which could be represented via nonlinear approximator $f(x)$. Then, there should be a $\tilde{f}(x)$ optimized via the following regularized loss function:

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \tilde{f}(x_i))^2 + \lambda\sum_{p=1}^{P}\hat{w}_p||\frac{\partial\tilde{f}}{\partial x_p}||_2$$

where $\hat{w}_p = \frac{1}{||\frac{\partial\hat{f}}{\partial x_p}||_2^\gamma}$. Then, the estimator $\tilde{f}(x)$ has oracle properties:

- 1 Consistency in variable selection: $\lim_n(||\frac{\partial\tilde{f}}{\partial x_p}||_2 \neq 0; ||\frac{\partial g}{\partial x_p}||_2 \neq 0) = 1$.

- 2 Point-wise asymptotic normality $\sqrt{n}(\tilde{f}(x) - f(x)) \rightarrow_d N(0, \sigma^2(x))$.

The proof is given in the Appendix.

5.2 Nonlinear Variable Selection via Soft Decision Tree with L_1 Regularization

Thus, the soft decision tree is differentiable, which means we can select variables via L_1 regularization, we can introduce the L_1-regularization into the log-likelihood function of soft decision tree:

$$\begin{aligned} Min: \ -L(y,x;\beta,c)+R_\lambda(f) = -\sum_{i=1}^{N} Log \sum_{node=1}^{S^2} \prod_{p=node} \sigma_p(x_i) P_{node}(y_i|x_i) \\ +\lambda \sum_{p=1}^{P} \hat{w}_p || \frac{\partial \tilde{f}}{\partial x_p} ||_2 \end{aligned} \tag{38}$$

where $||\frac{\partial \hat{f}}{\partial x_p}||_2 = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial \hat{f}}{\partial x_p}|_{x=x_i}\right)^2}$ and $\hat{w}_p = \frac{1}{||\frac{\partial \hat{f}}{\partial x_p}||_2^{\gamma}}$.

Based on the results in section 4.1, we could use the loss function instead of the log-likelihood function:

$$Min: \ Loss = \frac{1}{N}\sum_{i=1}^{N}(y_i - SDT(x_i))^2 + \lambda \sum_{p=1}^{P} \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial SDT}{\partial x_p}|_{x=x_i}\right)^2}}{\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial \widehat{SDT}}{\partial x_p}|_{x=x_i}\right)^2}\right)^{\gamma}} \tag{39}$$

where $SDT = SDT(x_i) = \prod_{p=node} \sigma_p(x_i)\mu_{node}$.

According to the Proposition 1, we can give the result 4 about consistent nonlinear variable selection:

Result 4: given the likelihood function of a soft decision tree with adaptive L_1 penalty, we have the following conclusions:

- **1** $P(f_p' = 0 | p \in A) \to 0$.
- **2** $\sqrt{n}\left(\tilde{f}(x) - f(x)\right) \sim N(0, J_\theta(x)^T \Sigma J_\theta(x))$.

6 Simulation Study

Based on the discussion in the previous sections, we can see that the L_1-regularized soft decision tree can identify the relevant variables consistently. In this section, we start the simulation experiment to show the performance of the L_1-regularized soft decision tree method. We choose Lasso as our baseline method. In the first study, we compare their variable selection results in 4 settings. In the second study, we compare their variable selection results in the partially linear regression settings.

6.1 Nonlinear Variable Selection in Regression Model

In the first study, we choose 4 data generating processes as follows:

- DGP1:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1 x_2 + u, u \sim N(0, 1).$$

- DGP2:

$$y = \beta_0 + \beta_1 sin(x_1) + \beta_2 x_2 \times exp(\beta_3 x_1 x_2) + \beta_4 x_3 + \beta_5 x_4 \times x_5 + u, u \sim N(0, 1).$$

- DGP3:

$$y = \beta_0 + \beta_1 sin(x_1) + \beta_2 x_2 + exp(\beta_3 x_1 x_2) \times u, u \sim N(0, 1).$$

- DGP4:

P($\beta \neq 0$|true)	**SDT**	**Lasso**	**Adaptive Lasso**
DGP 1	0.99	0.89	0.95
DGP 2	1	0.7	0.83
DGP 3	0.99	0.92	0.93
DGP 4	0.91	0.5	0.5

Table 1: Variable Selection for Nonlinear Regression ($p = 30$)

$$y = \beta_1(x_1^2 + x_2^2) \times exp(\beta_2 x_1^2 + x_2^2) \times +u, u \sim N(0, 1).$$

Precisely, we choose the input variable with 30 dimensions and only the first two or five dimensions are relevant. Then, the sample size is 5000. Finally, we decide the hyperparameters via 10-fold cross validation.

Table 1 shows the results of the variable selection based on soft decision tree, Lasso and Adaptive Lasso. We can see that to the simple DGP 1, SDT, Lasso and Adaptive Lasso work well in the variable selection. To the complicated and highly-nonlinear DGP 2, 3 and 4, SDT method outperforms the Lasso and Adaptive Lasso significantly.

6.2 Nonlinear Variable Selection in Treatment Effect Model

In the second study, we design the data generating process based on the generalized partially linear regression:

- $g_h(D, H) = \beta_{0,D} + \beta_{1,D} H_1^2 + \beta_{2,D} H_2 + \beta_{3,D} exp(H_3)$(Heterogeneous function in the outcomes)

- $g_c(D, C) = \alpha_{0,D} + \alpha_{1,D} \times exp(C_1) + \alpha_{2,D} \times C_2$(Confounding function in the outcomes)

- $$e(C,V) = \frac{1}{1+exp(-k(C,V))} \text{(Propensity Score)}$$

- $$k(C,V) = 1 + sin(C_1 \times \gamma_1 C_2) + \gamma_2 exp(V_1)$$

- $$y(D,H,C) = g_h(D,H) + g_c(D,C) + u,\ u \sim N(0,1) \text{(Outcomes)}$$

To each regression, only two variables are relevant, we choose the input dimension as 30 and the sample size is 5000.

Table 2 shows the results of the variable selection based on the Generalized Partially Linear Regression. We can see that SDT works better than the Lasso in variable selection.

7 Emperical Analysis: The effect of Unemployment Insurance Bonus on Unemployment Duration

In this example, we re-analyze the Pennsylvania Reemployment Bonus experiment which was conducted by the US Department of Labor in the 1980s to test the incentive effects of alternative compensation schemes for unemployment insurance (UI). This experiment has been previously studied by Bilias (2000) and Bilias and Koenker (2002). In these experiments, UI claimants were

P($\beta \neq 0$|true)	**SDT**	**Lasso**	**Adaptive Lasso**
y(D,H,C)	0.97	0.65	0.78
e(C,V)	0.98	0.74	0.82

Table 2: Variable Selection for Generalized Partially Linear Regression ($p = 30$)

Methods	**Heterogeneous**	**Confounding**
SDT	Jan, Feb, ... ,Dec	Site 1, Site 2, ... ,Site 12
Lasso	flagwag7, flagwag4, flagwag1	Site 1, Site 2, ... ,Site 12
Adaptive Lasso	aug, dob, Oct-89, mar, pjswcnt,wrwage8	Site 1, Site 2, ... ,Site 12

Table 3: Variable Selection for the Unemployment Insurance Bonus Data

randomly assigned either to a control group or one of five treatment groups. In the control group, the standard rules of the UI system applied. Individuals in the treatment groups were offered a cash bonus if they found a job within some pre-specified period (qualification period), provided that the job was retained for a specified duration. The treatments differed in the level of the bonus, the length of the qualification period, and whether the bonus was declining over time in the qualification period; see Bilias and Koenker (2002) for further details.

After cleaning the data and delete all the irrelevant features, the sample size is 14068. To the five treatment groups, we find that group 4 and group 6 are often combined into one treatment group since they represent very similar treatment. Thus, we do the same processing so that our result is comparable to the results from other papers. To the covariates, the input dimension is 70, including gender, age, sites, date and more.

Our main objective is to identify two groups of related attributes for heterogeneous effect and confounding effect respectively. We compare the Soft Decision Tree method with the Lasso and Adaptive Lasso. Then, the 10-fold validation method is used to decide the hyperparameters λ and γ.Our results are shown in Table 5.

Based on the selection result, we can see that:

- The location of the program is highly correlated with the treatment and outcome. Intuitively, it means that geographic information can distort the treatment effect.
- The month of the program decides the heterogeneous effect.
- The other factors are irrelevant since it is a random experiment.

Attractively, even though the experiment is applied as random as possible, there are still some confounders that can distort the randomness. Also, the heterogeneous effect could be very significant and provide that the time point of the program is very important to the outcomes.

More importantly, compared to SDT method, Lasso and Adaptive Lasso work well in selecting the confounders. But both of the linear methods cannot identify meaningful factors for the heterogeneous effect. Thus, given the unknown functions for outcomes y and propensity score $e(x)$, it is more reasonable to use our nonlinear variable selection method than linear methods.

8 Conclusions

To sum up, this paper proposed a L_1-regularized soft decision tree, a new method for identifying the heterogeneous factors and the confounders. We prove that the soft decision tree is equivalent to the kernel regression and it can approximate the unknown nonlinear function. Then, the L_1-regularized soft decision tree can identify the relevant variables consistently in the nonlinear framework and we find that it works in the simulation data and real data. Since L_1-regularized soft decision tree can identify the important variables based on weak assumptions, it can be applied in a variety of applications in the nonlinear regression and causal inference.

References

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Jared K. Lunceford and Marie Davidian. Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study. *Statistics in Medicine*, 23(19): 2937–2960, 2004. ISSN 02776715. doi: 10.1002/sim.1903.

Pearl, Judea. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.

Pearl, Judea and Mackenzie, Dana. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

Imbens, Guido W. and Rubin, Donald B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2017.

Donald B., Rubin. Estimating Causal Effects of Treatment in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Susan Athey. *The Impact of Machine Learning on Economics, The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press, 2019.

Susan Athey and Guido W. Imbens. Machine Learning Methods Economists Should Know About. *arXiv*, 2019. URL `http://arxiv.org/abs/1903.10075`.

Hal R. Varian. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2): 3–28, 2014. ISSN 08953309. doi: 10.1257/jep.28.2.3.

Adam Richardson, Thomas Mulder, and Tugru l Vehbi. Nowcasting New Zealand GDP using Machine learning Algorithms. *SSRN Electronic Journal*, 2018. doi: 10.2139/ssrn.3256578.

Greg Lewis and Vasilis Syrgkanis. Adversarial Generalized Method of Moments. *arXiv*, 2018. URL `http://arxiv.org/abs/1803.07164`.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A Flexible Approach for Counterfactual Prediction. In *The 34th International Conference on Machine Learning, Sydney, Australia*, 2016.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized Random Forests. *Annals of Statistics*, 47(2):1179–1203, 2019. ISSN 00905364. doi: 10.1214/18-AOS1709.

Mark J. van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2, 2006.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010. doi: 10.1214/09-AOAS285. URL `https://doi.org/10.1214/09-AOAS285`.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on Treatment Effects after Selection among High-Dimensional Controls. *Review of Economic Studies*, 81(2):608–650, 2013. ISSN 1467937X. doi: 10.1093/restud/rdt044.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, 28(2):29–50, 2014. ISSN 08953309. doi: 10.1257/jep.28.2.29.

Susan Athey and Guido Imbens. Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of the National Academy of Sciences of the United States of America*, 113(27):7353–7360, 2016. ISSN 1091-6490. doi: 10.1073/pnas.1510489113.

Ye Luo and Martin Spindler. Estimation and Inference of Treatment Effects with L_2-Boosting in High-Dimensional Settings. 2017. URL `http://arxiv.org/abs/1801.00364`.

Stefan Wager and Susan Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. ISSN 1537274X. doi: 10.1080/01621459.2017.1319839.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/Neyman Machine Learning of Treatment Effects. *American Economic Review*, 107(5):261–265, 2017. ISSN 00028282. doi: 10.1257/aer.p20171038.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/Debiased Machine Learning for Treatment and Causal Parameters. *The Econometrics Journal*, 21:C1–C68, 2018.

P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

Lester Mackey, Vasilis Syrgkanis, and Dias Zadik. Orthogonal Machine learning: Power and Limitations. *International Conference on Machine Learning*, 13:9112–9124, 2018.

Michael Jordan and Robert Jacob. Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Computation*, 6:181–214, 1994.

J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.*, 27(4):887–906, 12 1956. doi: 10.1214/aoms/1177728066. URL `https://doi.org/10.1214/aoms/1177728066`.

J. Pfanzagl. Regression shrinkage and selection via the lasso. *Journal of Statistical Planning and Inference*, 19:137–158, 1988.

Jiahua Chen. Consistency of the MLE under Mixture Models. *Statistical Science*, 32(1):47–63, 2017. ISSN 08834237. doi: 10.1214/16-STS578.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Jianqing Fan and Runze Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001. ISSN 0162-1459. doi: 10.1198/016214501753382273.

Hui Zou. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. ISSN 01621459. doi: 10.1198/016214506000000735.

Appendix 1: Proof for the Result 1

Previously, we optimize this function via searching the maximum of the complete likelihood function:

$$L_c(y,x;\beta,c,\hat{\mu},\Sigma)=\sum_{i=1}^{N}\sum_{node=1}^{S^2} z_{node} Log \prod_{p=node} \sigma_p(x_i;\beta_p,c_p) P_{node}(y_i;\hat{\mu}_{node},\Sigma_{node})$$

where z_{node} are implicit variables and the expectation of z is the posterior probability given all the data and parameters $E(z_{node})P(node|y,x,\theta)$. Thus we have $\sum_{node} z_{node} = 1$. Intuitively, this means that we optimize the log likelihood function via optimizing the lower bound $Q(y,x;\beta,c,\hat{\mu},\Sigma)$. Based on his conclusion, we could think about another formation of the lower bound:

$$L_q(y,x;\theta)=\sum_{i=1}^{N}\sum_{node=1}^{S^2}\prod_{p=node} g_p(x_i;\beta_p,c_p) LogP_{node}(y_i;\theta)=\sum_{i=1}^{N}\sum_{node=1}^{S^2}\alpha_{node} LogP_{node}(y_i;\theta)$$

Obviously, we have $L(y,x;\theta) \geq L_q(y,x;\theta)$. Precisely, if we consider $P_{node}(y;\theta) \sim N(\mu_{node},\sigma_{node})$, we have:

$$\begin{aligned}
L_q(y,x;\theta) &= \sum_{i=1}^{N}\sum_{node=1}^{S^2}\alpha_{node} LogP_{node}(y_i;\theta) \\
&= \sum_{i=1}^{N}\sum_{node=1}^{S^2}\alpha_{node} Log\frac{1}{\sqrt{2\pi\sigma_{node}^2}} e^{-\frac{(y_i-\mu_{node})^2}{\sigma_{node}^2}} \\
&= \sum_{i=1}^{N}\sum_{node=1}^{S^2}\alpha_{node}(-\frac{1}{2}(Log2\pi+Log\sigma_{node}^2)-\frac{(y_i-\mu_{node})^2}{\sigma_{node}^2}) \\
\rightarrow &-\sum_{i=1}^{N}\sum_{node=1}^{S^2}\alpha_{node}(\frac{1}{2}Log\sigma_{node}^2+\frac{(y_i-\mu_{node})^2}{\sigma_{node}^2})
\end{aligned} \tag{40}$$

If we assume $\sigma_{node} = \sigma$, we have:

$$
\begin{aligned}
L_q(y, x; \theta) &= -\sum_{i=1}^{N} \sum_{node=1}^{S^2} \alpha_{node} (\frac{1}{2} Log\sigma_{node}^2 + \frac{(y_i - \mu_{node})^2}{\sigma_{node}^2}) \\
&= -\sum_{i=1}^{N} \sum_{node=1}^{S^2} \alpha_{node} (\frac{1}{2} Log\sigma^2 + \frac{(y_i - \mu_{node})^2}{\sigma^2}) \\
&= -\sum_{i=1}^{N} \sum_{node=1}^{S^2} \alpha_{node} \frac{1}{2} Log\sigma^2 - \sum_{i=1}^{N} \sum_{node=1}^{S^2} \alpha_{node} \frac{(y_i - \mu_{node})^2}{\sigma^2} \\
&= -\sum_{i=1}^{N} \frac{1}{2} Log\sigma^2 \sum_{node=1}^{S^2} \alpha_{node} - \sum_{i=1}^{N} \frac{1}{\sigma^2} \sum_{node=1}^{S^2} \alpha_{node} (y_i - \mu_{node})^2 \\
&= -\frac{N}{2} Log\sigma^2 - \sum_{i=1}^{N} \frac{1}{\sigma^2} \sum_{node=1}^{S^2} \alpha_{node} (y_i - \mu_{node})^2 \\
\rightarrow &- \frac{N}{2} Log\sigma^2 - \frac{1}{\sigma^2} \sum_{i=1}^{N} (\sum_{node=1}^{S^2} \alpha_{node} y_i - \sum_{node=1}^{S^2} \alpha_{node} \mu_{node})^2 \\
&= -\frac{N}{2} Log\sigma^2 - \frac{1}{\sigma^2} \sum_{i=1}^{N} (y_i - \sum_{node=1}^{S^2} \alpha_{node} \mu_{node})^2 = L_{qq}(y, x; \theta)
\end{aligned}
\tag{41}
$$

We can see that the optimal x_i maximizing $\sum_i w_i x_i^2$ could also maximize $(\sum_i w_i x_i)^2$. Thus, we could maximize L_{pp} to get the maximum value of L_p. Furthermore, to likelihood function L_{pp}, we can see that $y \sim N(\sum_{node=1}^{S^2} \alpha_{node} \mu_{node}, \sigma^2)$. Thus, we could optimize all the parameters via minimizing the MSE loss function:

$$
Loss = \sum_{i=1}^{N} (y_i - \sum_{node=1}^{S^2} \alpha_{node} \mu_{node})^2
$$

Next, let's consider the estimator of σ^2, based on $L_{pp}(y, x; \theta)$, we have:

$$
\hat{\sigma}_{pp}^2 = \frac{1}{N} \sum_{i} (y_i - \sum_{node=1}^{S^2} \hat{\alpha}_{node} \hat{\mu}_{node})^2
$$

To $L_p(y, x; \theta)$, we have:

$$\hat{\sigma}_p^2 = \frac{1}{N} \sum_i \sum_{node=1}^{S^2} \hat{\alpha}_{node} (y_i - \hat{\mu}_{node})^2$$

To every sample point i, we have:

$$\begin{aligned}
&= \sum_{node=1}^{S^2} \hat{\alpha}_{node} (y_i - \hat{\mu}_{node})^2 \\
&= \sum_{node=1}^{S^2} \hat{\alpha}_{node} (y_i^2 - 2 y_i \hat{\mu}_{node} + \hat{\mu}_{node}^2) \\
&= \sum_{node=1}^{S^2} \hat{\alpha}_{node} y_i^2 - 2 \sum_{node=1}^{S^2} \hat{\alpha}_{node} y_i \hat{\mu}_{node} + \sum_{node=1}^{S^2} \hat{\alpha}_{node} \hat{\mu}_{node}^2 \\
&= y_i^2 - 2 y_i \sum_{node=1}^{S^2} \hat{\alpha}_{node} \hat{\mu}_{node} + \sum_{node=1}^{S^2} \hat{\alpha}_{node} \hat{\mu}_{node}^2
\end{aligned} \tag{42}$$

To L_{pp}, we have:

$$= y_i^2 - 2 y_i \sum_{node=1}^{S^2} \hat{\alpha}_{node} \hat{\mu}_{node} + (\sum_{node=1}^{S^2} \hat{\alpha}_{node} \hat{\mu}_{node})^2 \tag{43}$$

And,

$$\hat{\sigma}_p^2 - \sum_{node=1}^{S^2} \hat{\alpha}_{node} \hat{\mu}_{node}^2 = \hat{\sigma}_{pp}^2 - (\sum_{node=1}^{S^2} \hat{\alpha}_{node} \hat{\mu}_{node})^2$$

$$\hat{\sigma}_p^2 = \hat{\sigma}_{pp}^2 + (\sum_{node=1}^{S^2} \hat{\alpha}_{node} \hat{\mu}_{node}^2 - (\sum_{node=1}^{S^2} \hat{\alpha}_{node} \hat{\mu}_{node})^2))$$

Obviously, under the condition of equal variance, we can see that the estimator of variance based on L_{pp} is biased. But based on the formula above, we could get rid of this bias easily.

Appendix 2: Proof for Result 2

Now suppose we have a Gaussian Mixture Model with S components (kernels):

$$f(x) = \frac{\sum_{s=1}^{S} K(x; \beta_s, c_s)\mu_s}{\sum_{s=1}^{S} K(x; \beta_s, c_s)}$$

First of all, consider a simple case where $\beta_s = \beta$:

$$\begin{aligned} f(x) &= \frac{\sum_{s=1}^{S} K(x; \beta_s, c_s)\mu_s}{\sum_{s=1}^{S} K(x; \beta_s, c_s)} \\ &= \frac{\sum_{s=1}^{S} exp(-lg_2^S \beta(x - c_s)^2)\mu_s}{\sum_{s=1}^{S} exp(-lg_2^S \beta(x - c_s)^2)} \end{aligned} \tag{44}$$

Since $lg_2^S = log(S)/log(2)$, we have

$$\begin{aligned} f(x) &= \frac{\sum_{s=1}^{S} exp(-lg_2^S \beta(x - c_s)^2)\mu_s}{\sum_{s=1}^{S} exp(-lg_2^S \beta(x - c_s)^2)} \\ &= \frac{\sum_{s=1}^{S} exp(-log(S)/log(2)\beta(x - c_s)^2)\mu_s}{\sum_{s=1}^{S} exp(-log(S)/log(2)\beta(x - c_s)^2)} \\ &= \frac{\sum_{s=1}^{S} exp(-C * log(S)\beta(x - c_s)^2)\mu_s}{\sum_{s=1}^{S} exp(-C * log(S)\beta(x - c_s)^2)} \\ &= \frac{\sum_{s=1}^{S} exp(log(S^{(-C\beta(x-c_s)^2))})\mu_s}{\sum_{s=1}^{S} exp(log(S^{(-C\beta(x-c_s)^2))})} \\ &= \frac{\sum_{s=1}^{S} S^{(-C\beta(x-c_s)^2))}\mu_s}{\sum_{s=1}^{S} S^{(-C\beta(x-c_s)^2))}} \end{aligned} \tag{45}$$

where $C = 1/log(2)$.

We can see that the kernel is similar to a power kernel with S in the base.

Now consider the the bias:

$$\begin{aligned} Bias &= E|f_0(x) - f(x)| \\ &= E\left|f_0(x) - \frac{\sum_{s=1}^{S} S^{(-C\beta(x-c_s)^2))}\mu_s}{\sum_{s=1}^{S} S^{(-C\beta(x-c_s)^2))}}\right| \\ &= \frac{\sum_{s=1}^{S} S^{(-C\beta(x-c_s)^2))}}{\sum_{s=1}^{S} S^{(-C\beta(x-c_s)^2))}} E\left|f_0(x) - \mu_s\right| \end{aligned} \tag{46}$$

In the original soft decision tree, μ_s is treated as a parameter and estimated. In our paper, we introduce the definition **Honesty** from Wager and Athey [2018] such that $\mu_s = E(f_0(c_s)) = f_0(c_s)$. Precisely, we estimate μ_s via

$$\hat{\mu}_s = \hat{f}_0(c_s) = \frac{\sum_{i=1}^{N} K(x_i; \beta_s, c_s) y_i}{\sum_{i=1}^{N} K(x_i; \beta_s, c_s)} = f_0(c_s) + \varepsilon$$

Plug it into the previous equation, we have:

$$\begin{aligned} Bias &= \frac{\sum_{s=1}^{S} S^{(-C\beta(x-c_s)^2))}}{\sum_{s=1}^{S} S^{(-C\beta(x-c_s)^2))}} E\left|f_0(x) - \hat{\mu}_s\right| \\ &= \frac{\sum_{s=1}^{S} S^{(-C\beta(x-c_s)^2))}}{\sum_{s=1}^{S} S^{(-C\beta(x-c_s)^2))}} E\left|f_0(x) - f_0(c_s) - \varepsilon\right| \\ &= \frac{\sum_{s=1}^{S} S^{(-C\beta(x-c_s)^2))}}{\sum_{s=1}^{S} S^{(-C\beta(x-c_s)^2))}} E\left|f_0(x) - f_0(c_s)\right| + \varepsilon \end{aligned} \tag{47}$$

The next assumption we need is the so-called **Lipschitz continuity**. That is, we assume the target function $f_0(x)$ be changing slowly given finite support of x. That is, a Lipschitz continuous function is limited in how fast it can change:

$$|f_0(x) - f_0(x')| \leq D|x - x'|$$

Thus, our previous equation could be rewrite as:

$$\begin{aligned} Bias &= \frac{\sum_{s=1}^{S} S^{(-C\beta(x-c_s)^2))}}{\sum_{s=1}^{S} S^{(-C\beta(x-c_s)^2))}} E\left|f_0(x) - f_0(c_s)\right| + \varepsilon \\ &\leq D\frac{\sum_{s=1}^{S} S^{(-C\beta(x-c_s)^2))}}{\sum_{s=1}^{S} S^{(-C\beta(x-c_s)^2))}} \left|x - c_s\right| + \varepsilon \\ &= (1) + \varepsilon \end{aligned} \tag{48}$$

Let us focus on the part (1):

$$\begin{aligned} (1) &= D\frac{\sum_{s=1}^{S} S^{(-C\beta(x-c_s)^2))}}{\sum_{s=1}^{S} S^{(-C\beta(x-c_s)^2))}} \left|x - c_s\right| \\ &= D\frac{S^{(-C\beta(x-c_1)^2))}}{\sum_{s=1}^{S} S^{(-C\beta(x-c_s)^2))}} \left|x - c_1\right| + D\frac{S^{(-C\beta(x-c_2)^2))}}{\sum_{s=1}^{S} S^{(-C\beta(x-c_s)^2))}} \left|x - c_2\right| + ... + D\frac{S^{(-C\beta(x-c_s)^2))}}{\sum_{s=1}^{S} S^{(-C\beta(x-c_S)^2))}} \left|x - c_S\right| \\ &= D\frac{S^{(-C\beta d_1^2))}}{\sum_{s=1}^{S} S^{(-C\beta d_s^2))}} d_1 + D\frac{S^{(-C\beta d_2^2))}}{\sum_{s=1}^{S} S^{(-C\beta d_s^2))}} d_2 + ... + D\frac{S^{(-C\beta d_S^2))}}{\sum_{s=1}^{S} S^{(-C\beta d_s^2))}} d_S \end{aligned} \tag{49}$$

Let $d_m^2 : S^{(-C\beta d_m^2))} = 1/S \sum_{s=1}^{S} S^{(-C\beta d_s^2))}$, we have:

$$\begin{aligned} (1) &= D\frac{S^{(-C\beta d_1^2))}}{S * S^{(-C\beta d_m^2))}} d_1 + D\frac{S^{(-C\beta d_2^2))}}{S * S^{(-C\beta d_m^2))}} d_2 + ... + D\frac{S^{(-C\beta d_S^2))}}{S * S^{(-C\beta d_m^2))}} d_S \\ &= DS^{-(C\beta(d_1^2 - d_m^2)+1)} d_1 + DS^{-(C\beta(d_2^2 - d_m^2)+1)} d_2 + ... + DS^{-(C\beta(d_S^2 - d_m^2)+1)} d_S \end{aligned} \tag{50}$$

To each element $DS^{-(C\beta(d_s^2 - d_m^2)+1)} d_s$, we have:

$$Max_{d_s^2}\ DS^{-(C\beta(d_s^2 - d_m^2)+1)} d_s = DS^{-(C\beta(1/2 - d_m^2)+1)} 1/\sqrt{2}$$

where $d_s^{2*} = Argmax_{d_s^2}\ DS^{-(C\beta(d_s^2 - d_m^2)+1)} d_s = 1/\sqrt{2}$

Thus, we have:

$$\begin{aligned}(1) &= DS^{-(C\beta(d_1^2-d_m^2)+1)}d_1 + DS^{-(C\beta(d_2^2-d_m^2)+1)}d_2 + ... + DS^{-(C\beta(d_S^2-d_m^2)+1)}d_S \\ &\leq \frac{1}{\sqrt{2}}DS^{-C\beta(1/2-d_m^2)}\end{aligned} \tag{51}$$

Since $d_m \to 0$ when $S \to \infty$, it is guaranteed that:

$$(1) \leq \frac{1}{\sqrt{2}}DS^{-C\beta(1/2-d_m^2)} \to 0$$

Also, we need two more assumption:

1 $Var(c_s) \neq 0$

2 $|x|$ is finite.

To sum up, we have:

$$\begin{aligned}Bias &= (1) + \varepsilon \\ &\leq \frac{1}{\sqrt{2}}DS^{-C\beta(1/2-d_m^2)} + \varepsilon \to 0\end{aligned} \tag{52}$$

To the multi-variable case, suppose x is a p dimensional input vector. In Soft Decision Tree, we can choose a different kernel function:

$$S^{(-C(x-c_s)^T\Sigma_\beta(x-c_s)))}$$

where Σ_β is a weighted matrix. Suppose each dimension of x is independent, we can simplized kernel function as:

$$S^{(-C(x-c_s)^T\Sigma_\beta(x-c_s)))}$$

$$=S^{(-C\sum_{k=1}^{p}\beta_k(x_k-c_{ks})^2)}$$

The bias term should be:

$$\begin{aligned} Bias &= \frac{\sum_{s=1}^{S} S^{(-C\sum_{k=1}^{p}\beta_k(x_k-c_{ks})^2))}}{\sum_{s=1}^{S} S^{(-C\sum_{k=1}^{p}\beta_k(x_k-c_{ks})^2))}} E\left|f_0(x)-\hat{\mu}_s\right| \\ &= \frac{\sum_{s=1}^{S} S^{(-C\sum_{k=1}^{p}\beta_k(x_k-c_{ks})^2))}}{\sum_{s=1}^{S} S^{(-C\sum_{k=1}^{p}\beta_k(x_k-c_{ks})^2))}} E\left|f_0(x)-f_0(c_s)\right| + \varepsilon \end{aligned}$$

Considering the first term, according to the multivariate Lipchitz continuity $d(f_0(x), f_0(x')) \leq D \times d(x, x')$ where $d(x, y)$ is a metric, we choose the L_2 norm as the metric. The bias becomes:

$$\begin{aligned} Bias &= \frac{\sum_{s=1}^{S} S^{(-C\sum_{k=1}^{p}\beta_k(x_k-c_{ks})^2))}}{\sum_{s=1}^{S} S^{(-C\sum_{k=1}^{p}\beta_k(x_k-c_{ks})^2))}} E\left|f_0(x)-f_0(c_s)\right| \\ &= \frac{\sum_{s=1}^{S} S^{(-C\sum_{k=1}^{p}\beta_k(x_k-c_{ks})^2))}}{\sum_{s=1}^{S} S^{(-C\sum_{k=1}^{p}\beta_k(x_k-c_{ks})^2))}} d_{L_2}(x, c_s) \\ &= \frac{\sum_{s=1}^{S} S^{(-C\sum_{k=1}^{p}\beta_k(x_k-c_{ks})^2))}}{\sum_{s=1}^{S} S^{(-C\sum_{k=1}^{p}\beta_k(x_k-c_{ks})^2))}} \sqrt{\sum_{k=1}^{p}(x_k-c_{ks})^2} \\ &= \frac{\sum_{s=1}^{S} S^{(-C\sum_{k=1}^{p}\beta_k(d_{ks})^2))}}{\sum_{s=1}^{S} S^{(-C\sum_{k=1}^{p}\beta_k(d_{ks})^2))}} \sqrt{\sum_{k=1}^{p}(d_{ks})^2} \end{aligned}$$

Let $d_s^* : \sum_{k=1}^p \beta_k \times (d_s^*)^2 = \sum_{k=1}^p \beta_k (d_{ks})^2$, we have:

$$\begin{aligned} Bias &= \frac{\sum_{s=1}^S S^{(-C\sum_{k=1}^p \beta_k (d_{ks})^2))}}{\sum_{s=1}^S S^{(-C\sum_{k=1}^p \beta_k (d_{ks})^2))}} \sqrt{\sum_{k=1}^p (d_{ks})^2} \\ &\leq \frac{\sum_{s=1}^S S^{(-C\sum_{k=1}^p \beta_k (d_s^*)^2))}}{\sum_{s=1}^S S^{(-C\sum_{k=1}^p \beta_k (d_s^*)^2))}} \sqrt{\sum_{k=1}^p (d_s^*)^2} \\ &= \frac{\sum_{s=1}^S S^{(-C\sum_{k=1}^p \beta_k (d_s^*)^2))}}{\sum_{s=1}^S S^{(-C\sum_{k=1}^p \beta_k (d_s^*)^2))}} \sqrt{p} d_s^* \end{aligned} \tag{53}$$

Comparing the final result in equation (51) with the equation (46), we find that two results are the same. Thus, we can have the consistent result for the multiple case:

$$Bias \leq \frac{\sqrt{p}}{\sqrt{2}} D S^{-C|\beta|(1/2 - d_m^2)} + \varepsilon \to 0, \tag{54}$$

where $|\beta| = \sum_{k=1}^p \beta_k \sim O(p)$.

Appendix 3: Proof for Proposition 1 and Result 4

First, let's consider the estimator of L_2 norm of the first derivative $||\frac{\partial \hat{f}}{\partial x_p}||_2 = \sqrt{\frac{1}{N}\sum_{i=1}^N \left(\frac{\partial \hat{f}}{\partial x_p}|_{x=x_i}\right)^2}$. Based on the asymptotic normality of estimator of SDT's parameters $\sqrt{n}(\hat{\theta} - \theta) \sim N(0, \Sigma)$, we can have:

$$\sqrt{n}\left(\widehat{f_p'^2}|_i - f_p'^2|_i\right) \sim N(0, J^T(f_\theta''^2)\Sigma J(f_\theta''^2))$$

where $f_p'^2|_i = (\frac{\partial f}{\partial x_p}|_{x=x_i})^2$ and $J(f_\theta''^2)$ is the Jacobian matrix of $f_p'^2|_i$ to θ.

Thus, to $||\widehat{\frac{\partial f}{\partial x_p}}^2||_2$, we have:

$$\frac{\left(\widehat{f_p'^2|_i} - f_p'^2|_i\right)}{\sigma_i/\sqrt{n}} \sim N(0,1)$$

$$\widehat{f_p'^2|_i} - f_p'^2|_i \sim N(0, \sigma_i^2/n)$$

$$\sum_i \widehat{f_p'^2|_i} - \sum_i f_p'^2|_i \sim N(0, \sum_i \sigma_i^2/n)$$

$$\to n\left(\frac{1}{n}\sum_i \widehat{f_p'^2|_i} - \frac{1}{n}\sum_i f_p'^2|_i\right) \sim N(0, \sum_i \sigma_i^2/n)$$

$$\begin{cases} \sqrt{n}\sqrt{\frac{1}{n}\sum_i \widehat{f_p'^2|_i}} \sim N(0, \sqrt{\sum_i \sigma_i^2/n}) & if: \sqrt{\frac{1}{n}\sum_i f_p'^2|_i} = 0 \\ \sqrt{n}\left(\sqrt{\frac{1}{n}\sum_i \widehat{f_p'^2|_i}} - \sqrt{\frac{1}{n}\sum_i f_p'^2|_i}\right) \sim N(0, \frac{1}{2\sqrt{\frac{1}{n}\sum_i f_p'^2|_i}}\sum_i \sigma_i^2/n) & if: \sqrt{\frac{1}{n}\sum_i f_p'^2|_i} \neq 0 \end{cases}$$

which means $\frac{1}{n}\sum_i \widehat{f_p'^2|_i} - \frac{1}{n}\sum_i f_p'^2|_i \sim O_p(n^{-1})$.

Now we consider the penalized most likelihood function:

$$L_R = L(y, x; \theta) - \lambda R_\lambda(f') \tag{55}$$

where $L(y, x; \theta) = \sum_{i=1}^N Log \sum_{node=1}^{2^S} \alpha_{node}(x_i; \theta) P_{node}(y_i|x_i; \theta)$ and

$$R_\lambda(f') = \sum_{p=1}^P \frac{\sqrt{\frac{1}{N}\sum_{i=1}^N \left(\frac{\partial f}{\partial x_p}|_{x=x_i}\right)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^N \left(\frac{\partial \hat{f}}{\partial x_p}|_{x=x_i}\right)^2}^\gamma} = \sum_{p=1}^P R_\lambda^P(f')$$

By Taylor's expansion, we have:

$$\begin{aligned} L_R(y,x;\theta) &= L(y,x;\theta) - \lambda R_\lambda(f') \\ &= L(y,x;\theta_0) + (\theta-\theta_0)^T \frac{\partial L}{\partial \theta}|_{\theta_0} + 1/2(\theta-\theta_0)^T \frac{\partial L^2}{\partial\theta\partial\theta^T}|_{\theta_0}(\theta-\theta_0) - \lambda R_\lambda(f') \end{aligned} \tag{56}$$

Also, because of $\sqrt{n}(\hat{\theta}-\theta) \sim N(0,\Sigma)$, we set $\theta = \theta_0 + \frac{u}{\sqrt{n}}$. Then, we have:

$$\begin{aligned} -L_R(y,x;\theta) + L_R(y,x;\theta_0) &= -(\theta-\theta_0)^T \frac{\partial L}{\partial \theta}|_{\theta_0} - 1/2(\theta-\theta_0)^T \frac{\partial L^2}{\partial\theta\partial\theta^T}|_{\theta_0}(\theta-\theta_0) + \lambda(R_\lambda(f') - R_\lambda(f')_0) \\ &= Part(1) + Part(2) \end{aligned} \tag{57}$$

To $Part(1)$, we have:

$$\begin{aligned} Part(1) &= -(\theta-\theta_0)^T \frac{\partial L}{\partial \theta}|_{\theta_0} - 1/2(\theta-\theta_0)^T \frac{\partial L^2}{\partial\theta\partial\theta^T}|_{\theta_0}(\theta-\theta_0) \\ &= -(u/\sqrt{n})^T \frac{\partial L}{\partial \theta}|_{\theta_0} - 1/2(u/\sqrt{n})^T \frac{\partial L^2}{\partial\theta\partial\theta^T}|_{\theta_0}(u/\sqrt{n}) \end{aligned} \tag{58}$$

We could find that $\frac{\partial L}{\partial\theta}|_{\theta_0}$ is the score function of the log-likelihood function, and then $\frac{\partial L^2}{\partial\theta\partial\theta^T}|_{\theta_0}$ is the negative fisher information matrix. Thus, since $\frac{1}{\sqrt{n}}\frac{\partial L}{\partial\theta}|_{\theta_0} \to N(0, I(\theta))$ and $\frac{1}{n}\frac{\partial L^2}{\partial\theta\partial\theta^T}|_{\theta_0} \to I(\theta)$, we have:

$$\begin{aligned} Part(1) &= -(u/\sqrt{n})^T \frac{\partial L}{\partial \theta}|_{\theta_0} - 1/2(u/\sqrt{n})^T \frac{\partial L^2}{\partial\theta\partial\theta^T}|_{\theta_0}(u/\sqrt{n}) \\ &\to -u^T W + 1/2 u^T I(\theta) u \end{aligned} \tag{59}$$

where $W \sim N(0, I(\theta))$. Then, the $part(1)$ takes the maximum value when $u = I(\theta)^{-1}W \sim N(0, I(\theta)^{-1})$.

To $Part(2)$, we have:

$$
\begin{aligned}
&\lambda \left(R_\lambda^P(f') - R_\lambda^P(f')_0\right) \\
&= \lambda \left(\frac{\sqrt{\frac{1}{N}\sum_i \tilde{f}_p'^2|_i}}{\left(\sqrt{\frac{1}{N}\sum_i \hat{f}'^2_p|_i}\right)^\gamma} - \frac{\sqrt{\frac{1}{N}\sum_i f_p'^2|_i}}{\left(\sqrt{\frac{1}{N}\sum_i \hat{f}'^2_p|_i}\right)^\gamma} \right) \\
&= \lambda \left(\frac{\sqrt{\frac{1}{N}\sum_i \left(f_p'|_i + f''_{p,\theta}|_i^T \frac{u}{\sqrt{N}}\right)^2}}{\left(\sqrt{\frac{1}{N}\sum_i \hat{f}'^2_p|_i}\right)^\gamma} - \frac{\sqrt{\frac{1}{N}\sum_i f_p'^2|_i}}{\left(\sqrt{\frac{1}{N}\sum_i \hat{f}'^2_p|_i}\right)^\gamma} \right) \\
&= \frac{\lambda}{\left(\sqrt{\frac{1}{N}\sum_i \hat{f}'^2_p|_i}\right)^\gamma} \left(\sqrt{\frac{1}{N}\sum_i f_p'^2|_i + \frac{1}{N}\sum_i 2 f_p' f''_{p,\theta}|_i^T \frac{u}{\sqrt{N}} + \frac{1}{N}\sum_i (f''_{p,\theta}|_i^T \frac{u}{\sqrt{N}})^2} - \sqrt{\frac{1}{N}\sum_i f_p'^2|_i} \right)
\end{aligned}
\tag{60}
$$

Consider different cases, we have:

$$
= \begin{cases}
\frac{\lambda}{\sqrt{N}\left(\sqrt{\frac{1}{N}\sum_i \hat{f}'^2_p|_i}\right)^\gamma} \sqrt{\frac{1}{N}\sum_i (f''_{p,\theta}|_i^T u)^2} \to \frac{\lambda N^{\gamma/2}}{\sqrt{N}} O(1) \to \infty & if: \sqrt{\frac{1}{n}\sum_i f_p'^2|_i} = 0,\ \frac{1}{N}\sum_i (f''_{p,\theta}|_i^T u)^2 \neq \\
0 & if: \sqrt{\frac{1}{n}\sum_i f_p'^2|_i} = 0,\ \frac{1}{N}\sum_i (f''_{p,\theta}|_i^T u)^2 = \\
\frac{\lambda}{\sqrt{N}\left(\sqrt{\frac{1}{N}\sum_i \hat{f}'^2_p|_i}\right)^\gamma} \frac{\left(\frac{1}{N}\sum_i 2 f_p' f''_{p,\theta}|_i^T u + \frac{1}{N}\sum_i (f''_{p,\theta}|_i^T u)^2\right)}{\sqrt{\frac{1}{N}\sum_i \left(f_p'^2|_i\right)^*}} \to \frac{\lambda}{\sqrt{N}} O(1) \to 0 & if: \sqrt{\frac{1}{n}\sum_i f_p'^2|_i} \neq 0
\end{cases}
\tag{61}
$$

Now go back to $L_R(y, x; \theta) - L_R(y, x; \theta_0)$, we have:

$$\begin{aligned} -L_R(y,x;\theta) + L_R(y,x;\theta_0) &= -(\theta-\theta_0)^T \frac{\partial L}{\partial \theta}|_{\theta_0} - 1/2(\theta-\theta_0)^T \frac{\partial L^2}{\partial\theta\partial\theta^T}|_{\theta_0}(\theta-\theta_0) + \lambda(R_\lambda(f') - R_\lambda(f')_0) \\ &= -u^T W + 1/2 u^T I(\theta) u + Part(2) \end{aligned} \tag{62}$$

Let $V_n(u) = -L_R(y, x; \theta) + L_R(y, x; \theta_0)$, and we have:

$$V(u) = \begin{cases} -u^T W + 1/2 u^T I(\theta) u & if : \frac{1}{N}\sum_i (f''_{p,\theta}|_i^T u)^2 = 0, \forall p \notin A \\ \infty & otherwise \end{cases} \tag{63}$$

By slusky's theorem, we have $V_n(u) \to V(u)$. Thus, we have:

$$\hat{u} \to I(\theta)^{-1} W$$

$$\frac{1}{N}\sum_i (f''_{p,\theta}|_i^T \hat{u})^2 \to c\chi^2, \forall p \in A$$

$$\frac{1}{N}\sum_i (f''_{p,\theta}|_i^T \hat{u})^2 \to 0, \forall p \notin A$$

Intuitively, when $f'_p = 0$ is satisfied, variable should not be belong to the true variable set A. To the ture function $f(x_1, ..., x_{p-1})$, not only the first derivatie f'_p should be zero, but the derivative of f'_p to parameter θ: $f''_{p,\theta}$ should also be zero, which means $f(x)$ does not contain anything about x_p at even any parameters θ are not related to f'_p.

To sum up, we have approved the part of asymptotic normality:

$$\sqrt{n}\left(\tilde{f} - f\right) \sim N(0, J_\theta(x)^T \Sigma J_\theta(x))$$

under the penalized most likelihood estimation.

Next, we consider the consistency of variable selection.

From the asymptotic result, we can have:

$$P(\tilde{f}'_p \neq 0 | p \in A) \to 1$$

Thus, these variables should satisfy:

$$\frac{\partial L}{\partial \theta_s} = \lambda \sum_{p=1}^{P} \frac{1}{\left(\sqrt{\frac{1}{N}\sum_i \hat{f'}_p|_i^2}\right)^{\gamma}} \frac{\frac{1}{N}\sum_{i=1}^{N} \tilde{f}'_p|_i \tilde{f}''_{p,\theta_s}|_i}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} \tilde{f}'_p|_i^2}}$$

Since $\hat{f}'_p|_i$, $\hat{f}''_{p,\theta_s}|_i$ and $\hat{f}'_p|_i^2$ converge to zero at same convergence rate at $\sqrt{n}$, we have:

$$\frac{1}{N}\sum_{i=1}^{N} \hat{f}'_p|_i \hat{f}''_{p,\theta_s}|_i \sim \frac{1}{N}\sum_{i=1}^{N} \hat{f}'_p|_i^2 \sim O_p(N^{-1})$$

Obviously, the right hand side of the condition, we have:

$$\frac{\frac{1}{N}\sum_{i=1}^{N} \hat{f}'_p|_i \hat{f}''_{p,\theta_s}|_i}{\left(\frac{1}{N}\sum_{i=1}^{N} \hat{f}'_p|_i^2\right)^{1/2}} \sim \frac{1/N}{1/N^{1/2}} = \frac{1}{\sqrt{N}}$$

Combine with $\frac{\lambda N^{\gamma/2}}{\sqrt{N}} \to \infty$, we have:

$$\sum_{p=1}^{P} \frac{\lambda}{\left(\sqrt{\frac{1}{N}\sum_i \hat{f'}_p|_i^2}\right)^{\gamma}} \frac{\frac{1}{N}\sum_{i=1}^{N} \tilde{f}'_p|_i \tilde{f}''_{p,\theta_s}|_i}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} \tilde{f}'_p|_i^2}} \to \frac{\lambda N^{\gamma/2}}{\sqrt{N}} \to \infty$$

Since $\frac{\partial L}{\partial \theta_s} \to 0$, we have:

$$P(f'_p = 0 | p \in A) = P\left(-\frac{\partial L}{\partial \theta_s} = \sum_{p=1}^{P} \frac{\lambda}{\left(\sqrt{\frac{1}{N}\sum_i \hat{f'}_p|_i^2}\right)^{\gamma}} \frac{\frac{1}{N}\sum_{i=1}^{N} \tilde{f}'_p|_i \tilde{f}''_{p,\theta_s}|_i}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} \tilde{f}'_p|_i^2}}\right) \to 0$$

Thus, we have proved the oracle properties.