Proceedings of Machine Learning Research vol 75:1–14, 2018

Statistical Learnability of Generalized Additive Models based on Total Variation Regularization

Shin Matsushima SHIN_MATSUSHIMA@MIST.I.U-TOKYO.AC.JP
7-3-1, Hongo, Bunkyo, Tokyo, Japan

Abstract

A generalized additive model (GAM, Hastie and Tibshirani (1987)) is a nonparametric model by the sum of univariate functions with respect to each explanatory variable, i.e., $f(\boldsymbol{x}) = \sum f_j(x_j)$, where $x_j \in \mathbb{R}$ is j-th component of a sample $\boldsymbol{x} \in \mathbb{R}^p$. In this paper, we introduce the total variation (TV) of a function as a measure of the complexity of functions in $L^1_{\mathrm{c}}(\mathbb{R})$-space. Our analysis shows that a GAM based on TV-regularization exhibits a Rademacher complexity of $O(\sqrt{\log p / m})$, which is tight in terms of both m and p in the agnostic case of the classification problem. In result, we obtain generalization error bounds for finite samples according to work by Bartlett and Mendelson (2002).

Keywords: Generalized additive models, Generalization bound, Total variation, Rademacher complexity, Gaussian complexity

1. Introduction

In this paper, we focus on the learning problem of the following form of prediction functions:

$$f(\boldsymbol{x}) = \sum_{j \in [1,p]} f_j(x_j), \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^p$ denotes a sample and $x_j \in \mathbb{R}$ denotes the j-th explanatory variable for each $j \in [1,p] \triangleq \{j \in \mathbb{N} | 1 \leq j \leq p\}$. This was first proposed by Hastie and Tibshirani (1987) and is known as a *generalized additive model* (GAM). In this paper, we call $f_j(\cdot)$ a *weight function* and $f(\cdot)$ a *GAM predictor*. This not only includes linear predictors but also captures nonlinear relationships between explanatory variables and the targeted values. Although complex interactions or dependencies among explanatory values are not expressed, GAM predictors are expected to exhibit higher predictive performance when properly learned from a sufficiently large amount of data, at least in comparison with simple linear models. There has already been substantial work on data mining and statistics using GAMs (Guisan et al., 2002; Wood, 2006).

We first introduce the total variation (TV) of a function as a measure of complexity of functions in $L^1_{\mathrm{c}}(\mathbb{R})$-space. Here, $L^1_{\mathrm{c}}(\mathbb{R})$ denotes a space of functions with compact support in L^1-space on $\mathbb{R}$. Secondly, we introduce the sum of TV among all weight functions as a natural measure of complexity for GAM predictors:

Definition 1 *For $f \in L^1_{\mathrm{c}}(\mathbb{R})$, total variation of f denoted by* $\mathrm{TV}(f)$ *is defined as follows*

$$\mathrm{TV}(f(\cdot)) \triangleq \sup_{x:\mathbb{N}\to\mathbb{R},\,\text{increasing}} \sum_{n \in \mathbb{N}} |f(x(n)) - f(x(n+1))| \,.$$

Definition 2 *For $C \in \mathbb{R}_+$ and $p \in \mathbb{N}$, a set of TV-regularized GAM predictors denoted by $\mathrm{GAM}_p(C)$ is defined as follows:*

$$\mathrm{GAM}_p(C) \triangleq \left\{ f \in L^1_{\mathrm{c}}(\mathbb{R}^p) \,\middle|\, f(\boldsymbol{x}) = \sum_{j \in [1,p]} f_j(x_j), \sum_{j \in [1,p]} \mathrm{TV}(f_j) \leq C \right\}. \tag{2}$$

As we discuss in the next section, it has several desirable properties as a measure of complexity of functions under the framework of regularized empirical risk minimization. The main theorem of this paper (Theorem 6) states that the empirical Rademacher complexity of $\mathrm{GAM}_p(C)$ has an order of $O(C\sqrt{\log p/m})$. In result, we obtain generalization error bounds for finite samples according to the work by (Bartlett and Mendelson, 2002).

The main theorem is shown by analysis of the empirical Gaussian complexity without using concentration inequalities but with basic inequalities known in the field of stochastic process. This result implies that even discontinuous functions are learnable in GAM based on TV-regularization. For the paper to be self-contained, we state the definition of the empirical Rademacher complexity and the empirical Gaussian complexity below:

Definition 3 *For a set of functions $F \subset L^1(X)$ and $(\boldsymbol{x}_i)_{i \in [1,m]} \in X^m$, the empirical Gaussian complexity of F (with respect to $\boldsymbol{x}^m$) denoted by $G(F, (\boldsymbol{x}_i)_{i \in [1,m]})$ is defined as follows:*

$$G(F, (\boldsymbol{x}_i)_{i \in [1,m]}) \triangleq \frac{1}{m} \mathbb{E}_{\boldsymbol{\gamma}} \sup_{f \in F} \left| \sum_{i \in [1,m]} \gamma_i f(\boldsymbol{x}_i) \right|, \tag{3}$$

where each component of $\boldsymbol{\gamma} = (\gamma_i)_{i \in [1,m]}$ is an independent standard Gaussian random variable. Similarly, the empirical Rademacher complexity denoted by $R(F, (\boldsymbol{x}_i)_{i \in [1,m]})$ is defined as follows

$$R(F, (\boldsymbol{x}_i)_{i \in [1,m]}) \triangleq \frac{1}{m} \mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{f \in F} \left| \sum_{i \in [1,m]} \varepsilon_i f(\boldsymbol{x}_i) \right|. \tag{4}$$

where each component of $\boldsymbol{\varepsilon} = (\varepsilon_i)_{i \in [1,m]}$ is an independent Rademacher random variable[1].

In Section 2, we introduce several properties on TV-regularization. In Section 3, we show a technical lemma that will be used to prove the main theorem, which is formally stated and proven in Section 4. We conclude our paper with some discussion on related work and the tightness of our bound in Section 5.

2. Property of Total Variation (TV)

The formal expression of the TV-regularized empirical risk minimization for a GAM is given as follows:

$$\begin{aligned} &\mathrm{minimize}_{f, f_j} && \sum_{i \in [1,m]} \ell\left(f(\boldsymbol{x}_i), y_i\right) + \lambda \sum_{j \in [1,p]} \mathrm{TV}(f_j(\cdot)) \\ &\text{subject to} && f(\boldsymbol{x}) = \sum_{j \in [1,p]} f_j(x_j),\ f_j(\cdot) \in L^1_c(\mathbb{R}). \end{aligned} \tag{5}$$

1. A Rademacher random variable refers to a Bernoulli random variable of $\{\pm 1\}$ with the parameter $\frac{1}{2}$.

First, we state a type of compatibility of TV of weight functions to L^1-norm of a weight vector of linear predictors.

Property 1 *If ω is differentiable, then it holds that*

$$\mathrm{TV}(\omega) = \left\|\omega'\right\|_1 = \int_{\mathbb{R}} \left|\omega'(x)\right| \mathrm{d}x. \tag{6}$$

When we further restrict weight functions to be a form of $f_j(x_j) = w_j x_j [\![-M \le x_j \le M]\!]$[2] with a positive M, TV of f_j coincides with L^1-norm of $(w_j)_{j\in[1,p]}$ and hence the problem (5) is reduced to problems such as $L1$-logistic regression and LASSO (Tibshirani, 1996, 1997).

Second, we state that TV has a type of invariance and this property of TV leads to a desirable property of a solution of the problem (5).

Property 2 *Given a strictly monotone function $\varphi : \mathbb{R} \to \mathbb{R}$, it holds that*

$$\mathrm{TV}(\omega(\cdot)) = \mathrm{TV}(\omega \circ \varphi(\cdot))\,. \tag{7}$$

Proof $\mathrm{TV}(\omega)$ can be seen as a total variation of the signed measure $\omega(x)\mathrm{d}x$ and total variation of measures is invariant under bijective continuous transformations (Györfi, 2002). ∎

We note that this can be easily confirmed in case both ω and φ are differentiable as

$$\int_{\mathbb{R}} \left|\frac{\mathrm{d}\omega\circ\varphi(x)}{\mathrm{d}x}\right| \mathrm{d}x = \int_{\mathbb{R}} \left|\frac{\mathrm{d}\omega\circ\varphi(x)}{\mathrm{d}\varphi(x)}\frac{\mathrm{d}\varphi(x)}{\mathrm{d}x}\right| \mathrm{d}x = \int_{\mathbb{R}} \left|\frac{\mathrm{d}\omega\circ\varphi(x)}{\mathrm{d}\varphi(x)}\right| \frac{\mathrm{d}\varphi(x)}{\mathrm{d}x}\mathrm{d}x = \int_{\mathbb{R}} \left|\omega'(\varphi)\right| \mathrm{d}\varphi.$$

Property 3 *Given a strictly monotone function $\varphi_j : \mathbb{R} \to \mathbb{R}$ for each $j \in [1,p]$, consider modified samples*

$$(\hat{\boldsymbol{x}}_i, y_i)_{i\in[1,m]} \triangleq (\boldsymbol{\varphi}(\boldsymbol{x}_i), y_i)_{i\in[1,m]}, \tag{8}$$

where $\boldsymbol{\varphi}(\boldsymbol{x}) = (\varphi_j(x_j))_{j\in[1,p]}$. Then for any minimizer f^* in (5) with respect to $(\hat{\boldsymbol{x}}_i, y_i)_{i\in[1,m]}$, $f^* \circ \boldsymbol{\varphi}$ achieves the minimum of (5) with respect to $(\boldsymbol{x}_i, y_i)_{i\in[1,m]}$.

Proof For any f in (5) with respect to $(\hat{\boldsymbol{x}}_i, y_i)_{i\in[1,m]}$, $f \circ \boldsymbol{\varphi}$ yields the same value of objective function in (5) with respect to $(\boldsymbol{x}_i, y_i)_{i\in[1,m]}$. It can be seen from $f \circ \boldsymbol{\varphi} = \sum_{j\in[1,p]} f_j \circ \varphi_j$ and

$$\sum_{i\in[1,m]} \ell\left(f(\hat{\boldsymbol{x}}_i), y_i\right) + \lambda \sum_{j\in[1,p]} \mathrm{TV}(f_j(\cdot)) = \sum_{i\in[1,m]} \ell\left(f\circ\boldsymbol{\varphi}(\boldsymbol{x}_i), y_i\right) + \lambda \sum_{j\in[1,p]} \mathrm{TV}(f_j \circ \varphi_j(\cdot))\,.$$

∎

This property indicates that training based on TV-regularization is invariant under transformations of each explanatory variable such as scaling, shift and even nonlinear transformation by a monotone function. In the sense that we can obtain an optimal predictor among such transformation without any prior knowledge, this property is very important from a practical point of view.

Third, we see training based on TV-regularization is reduced to a minimization problem with a finite number of parameters.

2. $[\![\bullet]\!]$ is a function that returns 1 if $\bullet$ is true and 0 otherwise.

Property 4 *For a minimization problem defined in (5), consider the following minimization problem:*

$$\text{minimize}_{w_{j,s,t}} \sum_{i\in[1,m]} \ell\left(\sum_{(j,s,t)\in J} w_{j,s,t}\phi_{j,s,t}(\boldsymbol{x}_i), y_i\right) + 2\lambda \sum_{(j,s,t)\in J} |w_{j,s,t}|, \tag{9}$$

where $J = \{(j,s,t) \in [1,p] \times [1,m] \times [1,m] | s \le t\}$ *and*

$$\phi_{j,s,t}(\boldsymbol{x}) \triangleq \frac{1}{2}[\![x_{i(s)j} \le x_j < x_{i(t+1)j}]\!]. \tag{10}$$

Here, $i(t)$ *denotes the index of a sample at which the* j*-th explanatory variable is the* t*-th smallest among* $(x_{ij})_{i\in[1,m]}$*. Let* $(w^*_{j,s,t})_{(j,s,t)\in J}$ *be any minimizer of (9). Then, the minimum value of (5) is achieved by* f^* *where*

$$f^*(\boldsymbol{x}) = \sum_{(j,s,t)\in J} w^*_{j,s,t}\phi_{j,s,t}(\boldsymbol{x}). \tag{11}$$

Proof In (5), the first term of the objective function only depends on function values at observed samples. Meanwhile, while conditioning $v_{ij} = f_j(x_{ij})$ for $i \in [1,m]$, the problem of finding $f_j(\cdot)$ that minimizes TV and its minimum value can be analytically solved. The minimum is, for instance, achieved by the following function:

$$f_j(\cdot) = \sum_{t\in[1,m+1]} 2v_{i(t)j}\phi_{j,t-1,t}(\cdot). \tag{12}$$

Note that we defined exceptionally that $x_{i(0)j} \triangleq -M$ and $x_{i(m+1)j} \triangleq M$ with sufficiently large M and that $i(\cdot)$ implicitly depends on j. The optimality follows directly from the definition. Its total variation is expressed as

$$|v_{i(0)j}| + \sum_{t\in[1,m-1]} |v_{i(t)j} - v_{i(t+1)j}| + |v_{i(m)j}|. \tag{13}$$

Moreover, we can see that there exists $(w_{j,s,t})_{j\in J}$ which satisfies

$$\sum_{(j,s,t)\in J} w_{j,s,t}\phi_{j,s,t}(\cdot) = \sum_{j\in[1,p]}\sum_{t\in[1,m]} 2v_{i(t),j}\phi_{j,t-1,t}(\cdot) \quad \text{and} \tag{14}$$

$$2\sum_{(j,s,t)\in J} |w_{j,s,t}| = \sum_{j\in[1,p]}\left(|v_{i(0)j}| + \sum_{t\in[1,m-1]} |v_{i(t)j} - v_{i(t+1)j}| + |v_{i(n)j}|\right). \tag{15}$$

This technical lemma is shown in Lemma 5. Substituting these relationship, we obtain (9). ■

From this property, we can solve (9) to find the solution of (5). Therefore, when $\ell(\cdot, y)$ is convex for any $y \in Y$, it is boiled down to a convex minimization problem. Furthermore, as the second term has a separable structure, the coordinate-wise stationary condition guarantees the global solution when $\ell(\cdot, y)$ is also smooth. In this case, not only (5) is boiled down to a minimization problem with a finite number of parameters, but it can also be solved computationally efficiently.

Lastly, we state a property used in the proof of Lemma 5.

Property 5 *For any* $x < x'$*, let* $\phi_{x,x'}(\cdot) = \frac{1}{2}[\![x \le \cdot < x']\!]$*. Then* $\mathrm{TV}(\phi_{x,x'}) = 1$*.*

3. Technical Lemma

In this section, before the main theorem (Theorem 6) regarding the empirical Rademacher complexity of $\mathrm{GAM}_p(C)$, we prove a lemma (Lemma 4) used in its proof. Here, we define $A \circ B \triangleq \{(a,b) \in A \times B | a \leq b\}$ for $A, B \subset \mathbb{N}$.

Lemma 4 *Let $\gamma_1, \gamma_2, \ldots, \gamma_m$ and $x_1 \leq x_2 \leq \ldots \leq x_m$ be all real numbers. Then we have:*

$$2 \sup_{f \in \mathrm{GAM}_1(1)} \sum \gamma_i f(x_i) \leq \max_{i \in [0,m]} \Gamma_i - \min_{i \in [0,m]} \Gamma_i, \tag{16}$$

where $\Gamma_i \triangleq \sum_{j \in [1,i]} \gamma_j$.

Proof Using lemma 5, the value of $\sup_{f \in \mathrm{GAM}_1(1)} \sum \gamma_i f(x_i)$ has the following expression:

$$\sup_{f \in \mathrm{GAM}_1(1)} \sum \gamma_i f(x_i) = \sup \left\{ \sum_{(i,j) \in [1,m] \circ [1,m]} \Gamma_{ij} w_{ij} : \sum_{(i,j) \in [1,m] \circ [1,m]} |w_{ij}| \leq \frac{1}{2} \right\}, \tag{17}$$

where $\Gamma_{ij} \triangleq \sum_{i' \in [i,j]} \gamma_{i'}$. Obviously, this value is obtained by $\max_{(i,j) \in [1,m] \circ [1,m]} \frac{1}{2} |\Gamma_{ij}|$. Therefore, as we can easily see $\frac{1}{2} \max_{(i,j) \in [1,m] \circ [1,m]} |\Gamma_{ij}| \leq \frac{1}{2} \left(\max_{i \in [0,m]} \Gamma_i - \min_{i \in [0,m]} \Gamma_i \right)$, the claim of the lemma holds true. ∎

In what follows, we now state and prove lemma 5 used in the above.

Lemma 5 *Let $\gamma_1, \gamma_2, \ldots, \gamma_m$ and $x_1 \leq x_2 \leq \ldots \leq x_m$ be all real numbers. Then, the following statements are all equivalent:*

1. *there exists $f \in \mathrm{GAM}_1(1)$ such that $\sum_{i \in [1,m]} \gamma_i f(x_i) = \alpha$.*
2. *there exists $(v_i)_{i \in [1,m]}$ such that*

$$|v_1| + \sum_{i \in [1,m-1]} |v_i - v_{i+1}| + |v_m| \leq 1 \tag{18}$$

and $\sum_{i \in [1,m]} \gamma_i v_i = \alpha$.

3. *there exists $(w_{ij})_{(i,j) \in [1,m] \circ [1,m]}$ such that*

$$2 \sum_{(i,j) \in [1,m] \circ [1,m]} |w_{ij}| \leq 1, \quad \text{and} \tag{19}$$

$$\sum_{(i,j) \in [1,m] \circ [1,m]} \Gamma_{ij} w_{ij} = \alpha, \tag{20}$$

where $\Gamma_{ij} = \sum_{i' \in [i,j]} \gamma_{i'}$.

Proof $1 \to 2:$
For $f \in \mathrm{GAM}_1(1)$ such that $\sum_{i \in [1,m]} \gamma_i f(x_i) = \alpha$, set $v_i = f(x_i)$ for $i \in [1,m]$. From the definition of total variation, we can see

$$|f(x_1)| + \sum_{i \in [1,m-1]} |f(x_i) - f(x_{i+1})| + |f(x_m)| \leq \mathrm{TV}(f) \leq 1. \tag{21}$$

Then, $|v_1| + \sum_{i\in[1,m-1]} |v_i - v_{i+1}| + |v_m| \le 1$ and $\sum_{i\in[1,m]} \gamma_i v_i = \alpha$ are satisfied with $(v_i)_{i\in[1,m]}$.
$2 \to 3$:
We first prove that for any $(v_i)_{i\in[1,m]}$, there exists $(w_{ij})_{(i,j)\in[1,m]\circ[1,m]}$ such that

$$2\sum_{(i,j)\in[1,m]\circ[1,m]} |w_{ij}| = |v_1| + \sum_{i\in[1,m-1]} |v_i - v_{i+1}| + |v_m|, \quad \text{and} \tag{22}$$

$$\sum_{i\in[1,i']}\sum_{j\in[i',m]} w_{ij} = v_{i'}, \quad \forall i' \in [1,m] \tag{23}$$

by induction with respect to m. When $m = 1$, setting $w_{11} = v_1$ immediately gives (22) and (23). For $m \ge 2$, let $i^\star$ be the smallest index such that $v_{i^\star} = \max_i v_i$. For the simplicity of the notation, we set $v_0 = v_{m+1} = 0$. Then (22) can be written as

$$2\sum_{(i,j)\in[1,m]\circ[1,m]} |w_{ij}| = \sum_{i\in[0,m]} |v_i - v_{i+1}|. \tag{24}$$

By the inductive assumption, for $(v_i)_{i\in[1,m]\setminus i^\star}$, there exists $(w'_{ij})_{(i,j)\in([1,m]\setminus i^\star)\circ([1,m]\setminus i^\star)}$ such that

$$2\sum_{(i,j)\in[1,m]\circ[1,m]} |w'_{ij}| = |v_{i^\star+1} - v_{i^\star-1}| + \sum_{i\in[0,m-1]\setminus[i^\star-1,i^\star]} |v_i - v_{i+1}| \quad \text{and} \tag{25}$$

$$\sum_{i\in[1,i']\setminus i^\star}\sum_{j\in[i',m]\setminus i^\star} w'_{ij} = v_{i'} \quad i' \in [1,m]\setminus i^\star. \tag{26}$$

There are two possible cases: $v_{i^\star+1} < v_{i^\star-1}$ and $v_{i^\star+1} \ge v_{i^\star-1}$. In the case where $v_{i^\star+1} < v_{i^\star-1}$, we prove that $(w_{ij})_{(i,j)\in[1,m]\circ[1,m]}$ defined as follows satisfies (22) and (23):

$$w_{ij} = \begin{cases} w'_{ij} & (i,j) \in ([1,m]\setminus i^\star)\circ([1,m]\setminus[i^\star-1,i^\star]) \\ v_{i^\star} - v_{i^\star-1} & i = i^\star, j = i^\star \\ 0 & i = i^\star, j \in [i^\star+1, m] \\ 0 & i \in [1, i^\star-1], j = i^\star-1 \\ w'_{ii^\star-1} & i \in [1, i^\star-1], j = i^\star. \end{cases} \tag{27}$$

We can see (22) as follows:

$$\begin{aligned} 2\sum_{(i,j)\in[1,m]\circ[1,m]} |w_{ij}| &= 2\sum_{(i,j)\in([1,m]\setminus i^\star)\circ([1,m]\setminus i^\star)} |w'_{ij}| + 2|v_{i^\star} - v_{i^\star-1}| &(28)\\ &= |v_{i^\star+1} - v_{i^\star-1}| + \sum_{i\in[0,m]\setminus[i^\star-1,i^\star]} |v_i - v_{i+1}| + 2|v_{i^\star} - v_{i^\star-1}| &(29)\\ &= \sum_{i\in[0,m]\setminus[i^\star-1,i^\star]} |v_i - v_{i+1}| + v_{i^\star-1} - v_{i^\star+1} + 2v_{i^\star} - 2v_{i^\star-1} &(30)\\ &= \sum_{i\in[0,m]} |v_i - v_{i+1}|. &(31) \end{aligned}$$

As for (23), when $i' \in [1, i^\star - 1]$,

$$\sum_{i\in[1,i']}\sum_{j\in[i',m]} w_{ij} = \sum_{i\in[1,i']}\left(\sum_{j\in[i',m]\setminus[i^\star-1,i^\star]} w_{ij} + w_{ii^\star-1} + w_{ii^\star}\right) \tag{32}$$

$$= \sum_{i\in[1,i']}\left(\sum_{j\in[i',m]\setminus[i^\star-1,i^\star]} w'_{ij} + 0 + w'_{ii^\star-1}\right) \tag{33}$$

$$= \sum_{i\in[1,i']\setminus i^\star}\left(\sum_{j\in[i',m]\setminus i^\star} w'_{ij}\right) \tag{34}$$

$$= v_{i'}. \tag{35}$$

When $i' = i^\star$,

$$\begin{aligned}
\sum_{i\in[1,i^\star]}\sum_{j\in[i^\star,m]} w_{ij} &= \sum_{i\in[1,i^\star]}\left(\sum_{j\in[i^\star+1,m]} w_{ij} + w_{ii^\star}\right)\\
&= \sum_{i\in[1,i^\star-1]}\left(\sum_{j\in[i^\star+1,m]} w_{ij} + w_{ii^\star}\right) + \sum_{j\in[i^\star+1,m]} w_{i^\star j} + w_{i^\star i^\star}\\
&= \sum_{i\in[1,i^\star-1]}\left(\sum_{j\in[i^\star+1,m]} w'_{ij} + w'_{ii^\star-1}\right) + 0 + v_{i^\star} - v_{i^\star-1}\\
&= \sum_{i\in[1,i^\star-1]\setminus i^\star}\ \sum_{j\in[i^\star-1,m]\setminus i^\star} w'_{ij} + v_{i^\star} - v_{i^\star-1}\\
&= v_{i^\star-1} + v_{i^\star} - v_{i^\star-1} = v_{i^\star}.
\end{aligned}$$

When $i' \in [i^\star + 1, m]$,

$$\sum_{i\in[1,i']}\sum_{j\in[i',m]} w_{ij} = \sum_{i\in[1,i']\setminus i^\star}\ \sum_{j\in[i',m]} w_{ij} + \sum_{j\in[i',m]} w_{i^\star j} \tag{36}$$

$$= \sum_{i\in[1,i']\setminus i^\star}\ \sum_{j\in[i',m]\setminus[i^\star-1,i^\star]} w_{ij} \tag{37}$$

$$= \sum_{i\in[1,i']\setminus i^\star}\ \sum_{j\in[i',m]\setminus[i^\star-1,i^\star]} w'_{ij} \tag{38}$$

$$= v_{i'}. \tag{39}$$

Therefore, (23) holds. In the case where $v_{i^\star+1} \geq v_{i^\star-1}$, set $(w_{ij})_{(i,j)\in[1,m]\circ[1,m]}$ as follows:

$$w_{ij} = \begin{cases} w'_{ij} & (i,j) \in ([1,m] \setminus [i^\star, i^\star+1]) \circ ([1,m] \setminus i^\star) \\ v_{i^\star} - v_{i^\star+1} & i = i^\star, j = i^\star \\ w'_{i^\star+1j} & i = i^\star, j \in [i^\star+1, m], \\ 0 & i = i^\star+1, j \in [i^\star+1, m], \\ 0 & i \in [1, i^\star-1], j = i^\star. \end{cases} \tag{40}$$

As for (22), a similar argument as above holds as follows

$$2\sum_{(i,j)\in[1,m]\circ[1,m]}|w_{ij}| = |v_{i^\star+1}-v_{i^\star-1}| + \sum_{i\in[0,m]\setminus[i^\star-1,i^\star]}|v_i - v_{i+1}| + 2(v_{i^\star}-v_{i^\star+1}) \tag{41}$$

$$= \sum_{i\in[0,m]}|v_i - v_{i+1}|. \tag{42}$$

Also for (23), in case $i' \in [1, i^\star - 1]$,

$$\sum_{i\in[1,i']}\sum_{j\in[i',m]} w_{ij} = \sum_{i\in[1,i']}\sum_{j\in[i',m]\setminus i^\star} w_{ij} = \sum_{i\in[1,i']\setminus[i^\star,i^\star+1]}\sum_{j\in[i',m]\setminus i^\star} w_{ij} = v_{i'}. \tag{43}$$

In case $i' = i^\star$,

$$\begin{aligned}\sum_{i\in[1,i^\star]}\sum_{j\in[i^\star,m]} w_{ij} &= \sum_{i\in[1,i^\star-1]}\left(\sum_{j\in[i^\star+1,m]} w_{ij} + w_{ii^\star}\right) + \sum_{j\in[i^\star+1,m]} w_{i^\star j} + w_{i^\star i^\star} \\ &= \sum_{i\in[1,i^\star-1]}\left(\sum_{j\in[i^\star+1,m]} w'_{ij} + 0\right) + \sum_{j\in i^\star+[1,m]} w'_{i^\star+1j} + v_{i^\star} - v_{i^\star+1} \\ &= \sum_{i\in[1,i^\star+1]\setminus i^\star}\sum_{j\in[i^\star+1,m]\setminus i^\star} w'_{ij} + v_{i^\star} - v_{i^\star+1} \\ &= v_{i^\star+1} + v_{i^\star} - v_{i^\star+1} = v_{i^\star}.\end{aligned}$$

Finally, in case $i' \in [i^\star + 1, m]$,

$$\sum_{i\in[1,i']}\sum_{j\in[i',m]} w_{ij} = \sum_{j\in[i',m]}\left(\sum_{i\in[1,i']\setminus[i^\star,i^\star+1]} w_{ij} + w_{i^\star j} + w_{i^\star+1j}\right) \tag{44}$$

$$= \sum_{j\in[i',m]}\left(\sum_{i\in[1,i']\setminus[i^\star,i^\star+1]} w'_{ij} + w'_{i^\star+1j} + 0\right) \tag{45}$$

$$= \sum_{j\in[i',m]\setminus i^\star}\left(\sum_{i\in[1,i']\setminus i^\star} w'_{ij}\right) \tag{46}$$

$$= v_{i'}. \tag{47}$$

Therefore, (22) and (23) hold for any $m \in \mathbb{N}$. We then show such $(w_{ij})_{(i,j)\in[1,m]\circ[1,m]}$ satisfies (19) and (20) given that 2. holds. (19) can be immediately seen by (18) and (22). For (20), it holds from $\sum_{i\in[1,m]}\gamma_i v_i = \alpha$ and the following relation:

$$\sum_{(i,j)\in[1,m]\circ[1,m]}\Gamma_{ij}w_{ij} = \sum_{(i,j)\in[1,m]\circ[1,m]}\sum_{i'\in[i,j]}\gamma_{i'}w_{ij} = \sum_{i'\in[1,m]}\sum_{(i,j)\in[1,i']\times[i',m]}\gamma_{i'}w_{ij} = \sum_{i'\in[1,m]}\gamma_{i'}v_{i'}.$$

$3 \to 1$:
For such $(w_{ij})_{(i,j)\in[1,m]\circ[1,m]}$, set $f(\cdot)$ as $\sum_{(i,j)\in[1,m]\circ[1,m]} 2w_{ij}\phi_{i,j}(\cdot)$, where $\phi_{i,j}(\cdot) = \frac{1}{2}[\![x_i \le \cdot \le$

$x_j]\!]$. As $\mathrm{TV}(\phi_{i,j}) = 1$,

$$\mathrm{TV}(f) \leq \sum_{(i,j)\in[1,m]\circ[1,m]} 2|w_{ij}|\mathrm{TV}(\phi_{i,j}(\cdot)) \leq \sum_{(i,j)\in[1,m]\circ[1,m]} 2|w_{ij}| \leq 1, \tag{48}$$

which means $f \in \mathrm{GAM}_1(1)$. On the other hand, because $f(x_{i'}) = \sum_{(i,j)\in[1,i']\times[i',m]} w_{ij}$, we see

$$\sum_{i'\in[1,m]} \gamma_{i'} f(x_{i'}) = \sum_{i'\in[1,m]} \gamma_{i'} \sum_{(i,j)\in[1,i']\times[i',m]} w_{ij} = \sum_{(i,j)\in[1,m]\circ[1,m]} \Gamma_{ij} w_{ij} = \alpha. \tag{49}$$

■

4. Main Result

In this section, we state our main result on the empirical Rademacher complexity and the corollary on generalization bounds.

Theorem 6 *Let* $\mathrm{GAM}_{p,\ell}(C) \triangleq \{(\boldsymbol{x}, y) \mapsto \ell(f(\boldsymbol{x}), y) | f \in \mathrm{GAM}_p(C)\} \subset L^1(\mathbb{R}^p \times Y)$ *for a loss function* $\ell : \mathbb{R} \times Y \to \mathbb{R}_+$ *in which* $\ell(\cdot, y)$ *is* ρ*-Lipschitz for any* $y \in Y$*. Then, for any* $(\boldsymbol{x}_i, y_i)_{i\in[1,m]}$ *and* $d > 2$*, it holds that*

$$G(\mathrm{GAM}_{p,\ell}(C), (\boldsymbol{x}_i, y_i)_{i\in[1,m]}) \leq \sqrt{\frac{2}{\pi}}\rho C\sqrt{\frac{5\lceil \log p\rceil}{m}}, \tag{50}$$

and

$$R(\mathrm{GAM}_{p,\ell}(C), (\boldsymbol{x}_i, y_i)_{i\in[1,m]}) \leq \rho C\sqrt{\frac{5\lceil \log p\rceil}{m}}. \tag{51}$$

Proof We can easily see that (50) implies (51) from the following inequality:

$$\mathbb{E}_{\boldsymbol{\gamma}} \sup_{f\in F} \sum \gamma_i f(x_i) = \mathbb{E}_{\boldsymbol{\varepsilon}}\mathbb{E}_{\boldsymbol{\gamma}} \sup_{f\in F} \sum \varepsilon_i |\gamma_i| f(x_i) \tag{52}$$

$$\geq \mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{f\in F} \sum \varepsilon_i \mathbb{E}_{\gamma_i} |\gamma_i| f(x_i) \tag{53}$$

$$= \mathbb{E}_{\boldsymbol{\varepsilon}}\mathbb{E}_{\gamma_1} |\gamma_1| \sup_{f\in F} \sum \varepsilon_i f(x_i), \tag{54}$$

and $\mathbb{E}_{\gamma_1}|\gamma_1| = \sqrt{\frac{2}{\pi}}$. Therefore, from properties of the Rademacher complexity (Shalev-Shwartz and Ben-David, 2014, Lemma 26.6, Lemma 26.9), it is sufficient to prove that

$$G(\mathrm{GAM}_p(1), (\boldsymbol{x}_i)_{i\in[1,m]}) \leq \sqrt{\frac{5\lceil \log p\rceil}{m}}\sqrt{\frac{2}{\pi}}. \tag{55}$$

First, for any $r \geq 1$, we can see that

$$\sup_{f \in \mathrm{GAM}_p(1)} \sum_{i \in [1,m]} \gamma_i f(\boldsymbol{x}_i) = \sup \left\{ \sum_{j \in [1,p]} c_j \sup_{f \in \mathrm{GAM}_1(1)} \sum_i \gamma_i f_j(x_{ij}) : c_j \geq 0, \sum_j c_j \leq 1 \right\} \tag{56}$$

$$= \max_{j \in [1,p]} \sup_{f_j \in \mathrm{GAM}_1(1)} \sum_i \gamma_i f_j(x_{ij}) \tag{57}$$

$$\leq \left(\sum_{j \in [1,p]} \left(\sup_{f_j \in \mathrm{GAM}_1(1)} \sum_i \gamma_i f_j(x_{ij}) \right)^r \right)^{\frac{1}{r}}. \tag{58}$$

From Lemma 4,

$$2 \sup_{f \in \mathrm{GAM}_1(1)} \sum \gamma_i f(x_i) \leq \max_{i \in [0,m]} \Gamma_i - \min_{i \in [0,m]} \Gamma_i, \tag{59}$$

where $\Gamma_i = \sum_{j \in [1,i]} \gamma_j$. Therefore, $\sup_{f \in \mathrm{GAM}_1(1)} \sum \gamma_i f(x_i) > t$ implies that $\max_{i \in [0,m]} \Gamma_i - \min_{i \in [0,m]} \Gamma_i > 2t$, which then implies at least either $\max_{i \in [0,m]} \Gamma_i > t$ or $\min_{i \in [0,m]} \Gamma_i < -t$ holds. Therefore, for any $t > 0$, it holds that

$$\mathbb{P} \left\{ \sup_{f \in \mathrm{GAM}_1(1)} \sum_{i \in [1,m]} \gamma_i f(x_i) > t \right\} \leq \mathbb{P} \left\{ \max_{i \in [0,m]} \Gamma_i > t \right\} + \mathbb{P} \left\{ \min_{i \in [0,m]} \Gamma_i < -t \right\} \tag{60}$$

$$= 2\mathbb{P} \left\{ \max_{i \in [0,m]} \Gamma_i > t \right\}. \tag{61}$$

From Levy inequality (Ledoux and Talagrand, 1991), we see

$$\mathbb{P} \left\{ \max_{i \in [0,m]} \Gamma_i > t \right\} \leq 2\mathbb{P} \left\{ \Gamma_m > t \right\} = 2 \int_{s=t}^{\infty} \frac{1}{\sqrt{2\pi m}} e^{-\frac{s^2}{2m}} \mathrm{d}s. \tag{62}$$

Therefore, for any j,

$$\mathbb{E}_{\boldsymbol{\gamma}}\left(\sup_{f_j\in\mathrm{GAM}_1(1)}\sum_{i\in[1,m]}\gamma_i f_j(x_{ij})\right)^r = \int_{t=0}^{\infty}\mathbb{P}\left\{\left|\sup_{f_j\in\mathrm{GAM}_1(1)}\sum_{i\in[1,m]}\gamma_i f_j(x_{ij})\right|^r > t\right\}\mathrm{d}t \tag{63}$$

$$= \int_{t=0}^{\infty}\mathbb{P}\left\{\sup_{f_j\in\mathrm{GAM}_1(1)}\sum_{i\in[1,m]}\gamma_i f_j(x_{ij}) > t^{\frac{1}{r}}\right\}\mathrm{d}t \tag{64}$$

$$\leq \int_{t=0}^{\infty}2\mathbb{P}\left\{\max_{i\in[0,m]}\Gamma_i > t^{\frac{1}{r}}\right\}\mathrm{d}t \tag{65}$$

$$\leq \int_{t=0}^{\infty}4\int_{s=t^{\frac{1}{r}}}^{\infty}\frac{1}{\sqrt{2\pi m}}e^{-\frac{s^2}{2m}}\mathrm{d}s\mathrm{d}t \tag{66}$$

$$= 4\int_{s=0}^{\infty}\int_{t=0}^{s^r}\frac{1}{\sqrt{2\pi m}}e^{-\frac{s^2}{2m}}\mathrm{d}s\mathrm{d}t \tag{67}$$

$$= 4\int_{s=0}^{\infty}\frac{s^r}{\sqrt{2\pi m}}e^{-\frac{s^2}{2m}}\mathrm{d}s \tag{68}$$

$$= 2\mathbb{E}_{s\sim\mathrm{Normal}(0,m)}[|s|^r] \tag{69}$$

$$= 2(2m)^{\frac{r}{2}}\frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{\pi}}. \tag{70}$$

Finally, we see

$$\mathbb{E}_{\boldsymbol{\gamma}}\sup_{f\in\mathrm{GAM}_p(1)}\sum_{i\in[1,m]}\gamma_i f(\boldsymbol{x}_i) \leq \mathbb{E}_{\boldsymbol{\gamma}}\left(\sum_{j\in[1,p]}\left(\sup_{f_j\in\mathrm{GAM}_1(1)}\sum_{i\in[1,m]}\gamma_i f_j(x_{ij})\right)^r\right)^{\frac{1}{r}} \tag{71}$$

$$\leq \left(\mathbb{E}_{\boldsymbol{\gamma}}\sum_{j\in[1,p]}\left(\sup_{f_j\in\mathrm{GAM}_1(1)}\sum_{i\in[1,m]}\gamma_i f_j(x_{ij})\right)^r\right)^{\frac{1}{r}} \tag{72}$$

$$\leq \left(p\cdot 2(2m)^{\frac{r}{2}}\frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{\pi}}\right)^{\frac{1}{r}} \tag{73}$$

$$< \sqrt{2m}\left(\frac{2p}{\sqrt{\pi}}\sqrt{2\pi}\left(\frac{s}{2e}\right)^{\frac{s}{2}}e^{\frac{1}{6s}}\right)^{\frac{1}{1+s}}. \tag{74}$$

We set $r = 1+s$ and used $\Gamma(1+\frac{s}{2}) < \sqrt{2\pi}\left(\frac{s}{2e}\right)^{\frac{s}{2}}e^{\frac{1}{6s}}$ in the last inequality. Setting $s = 2\lceil\log p\rceil$,

$$\sqrt{2m}\left(2\sqrt{2}e^{\frac{s}{2}}\left(\frac{s}{2e}\right)^{\frac{s}{2}}e^{\frac{1}{6s}}\right)^{\frac{1}{1+s}} = \sqrt{2m}\sqrt{\frac{s}{2}}\left(2\sqrt{2}\sqrt{\frac{2}{s}}e^{\frac{1}{6s}}\right)^{\frac{1}{1+s}} \tag{75}$$

$$= \sqrt{2m\lceil\log p\rceil}\left(\frac{4}{\sqrt{s}}e^{\frac{1}{6s}}\right)^{\frac{1}{1+s}} \tag{76}$$

$$< \sqrt{2m\lceil\log p\rceil}\sqrt{\frac{5}{\pi}}. \tag{77}$$

(77) holds when $s \geq 4$ because $\left(\frac{4}{\sqrt{s}} e^{\frac{1}{6s}}\right)^{\frac{1}{1+s}}$ is maximized at $s = 4$ in the range $s \geq 4$ and $\left(2e^{\frac{1}{24}}\right)^{\frac{1}{5}}$ is less than $\sqrt{\frac{5}{\pi}}$ [3]. ■

Lastly, we state the generalization bound that can be derived directly from the result in (Shalev-Shwartz and Ben-David, 2014, Theorem 26.5).

Corollary 7 *Assume that $(\boldsymbol{x}, y)$ and $(\boldsymbol{x}_i, y_i)_{i\in[1,m]}$ are i.i.d. random variables on $\mathbb{R}^p \times Y$ and $\ell(\cdot, y)$ is ρ-Lipschitz and bounded by $c > 0$ for any $y \in Y$. Then, the following statements holds true for $p > 2$ and $\delta > 0$:*

1. *For any $f \in \mathrm{GAM}_p(C)$,*

$$\mathbb{P}_{(\boldsymbol{x}_i,y_i)_{i\in[1,m]}}\left\{\mathbb{E}_{\boldsymbol{x},y}\ell(f(\boldsymbol{x}), y) \leq \frac{1}{m}\sum_{i\in[1,m]} \ell(f(\boldsymbol{x}_i), y_i)) + \rho C\sqrt{\frac{5\lceil\log p\rceil}{m}} + c\sqrt{\frac{2\log(2/\delta)}{m}}\right\} \geq 1 - \delta.$$

2. *For $\ell^* = \inf_{f\in\mathrm{GAM}_p(C)} \mathbb{E}_{\boldsymbol{x},y}\ell(f(\boldsymbol{x}), y)$ and $\hat{f} = \mathrm{argmin}_{f\in\mathrm{GAM}_p(C)} \frac{1}{m}\sum_{i\in[1,m]} \ell(f(\boldsymbol{x}_i), y_i)$,*

$$\mathbb{P}_{(\boldsymbol{x}_i,y_i)_{i\in[1,m]}}\left\{\mathbb{E}_{\boldsymbol{x},y}\ell(\hat{f}(\boldsymbol{x}), y) \leq \ell^* + \rho C\sqrt{\frac{5\lceil\log p\rceil}{m}} + 5c\sqrt{\frac{2\log(2/\delta)}{m}}\right\} \geq 1 - \delta.$$

5. Discussion

5.1. Related Work

Probably, the closest result to ours is work by Cortes et al. (2010), in which the authors studied the Rademacher complexity of the following hypothesis class in the context of multiple kernel learning:

$$H_p^1 = \left\{ f(\boldsymbol{x}) = \sum_{j\in[1,p]} \mu_j \langle\omega, \Phi_j(\boldsymbol{x})\rangle_j \,\middle|\, \sum_{j\in[1,p]} \mu_j \|\omega\|_{\mathcal{H}_j} \leq 1, \sum_{j\in[1,p]} \mu_j \leq 1, \mu_j \geq 0 \right\},$$

Here, $\Phi_j(\boldsymbol{x}) \triangleq K_j(\boldsymbol{x}, \cdot)$ in which K_j is the reproducing kernel of RKHS $\mathcal{H}_j$ and $\|\cdot\|_{\mathcal{H}_j}$ and $\langle\cdot,\cdot\rangle_j$ denote its norm and inner product. The authors have shown that $R(H_p^1, (\boldsymbol{x}_i)_{i\in[1,m]})$ is an order of $O(\sqrt{\log p/m})$. When we restrict K_j to be dependent on the j-th explanatory variable only, H_p^1 becomes also a class of GAM predictors. However, each weight function has to be represented as $\mu_j \langle\omega, \Phi_j(x_j)\rangle_j$ by the same ω among all j and it is rather restrictive from the viewpoint of GAM predictors.

When we restrict $\|\boldsymbol{x}_i\|_\infty \leq 1$, Kakade et al. (2009) showed that $R(F_W, (\boldsymbol{x}_i)_{\in[1,m]})$ is an order of $O(\sqrt{\log p/m})$, where $F_W = \left\{\boldsymbol{x} \mapsto \sum_{j\in[1,p]} w_j x_j \,\middle|\, (w_j)_{j\in[1,p]} \in W\right\}$ and $W = \{\boldsymbol{w} \in \mathbb{R}^p | \|\boldsymbol{w}\|_1 \leq 1\}$. As it is easy to see $F_W \subset \mathrm{GAM}_p(2)$, we can view our result as an extension of their result to non-linear GAM predictors.

3. in case of $p = 2$, setting $r = 3$ in (73) yields a similar bound with $\sqrt{\frac{6}{\pi}}$ instead of $\sqrt{\frac{5}{\pi}}$

5.2. Tightness

We consider the result of Theorem 6 in the context of the classification problem, in which $Y = \{\pm 1\}$ and $X = \{\pm 1\}^p \subset \mathbb{R}^p$. Then, $J_p = \{\boldsymbol{x} \mapsto \pm\operatorname{sign}(x_j) | j \in [1,p]\} \subset \mathrm{GAM}_p(2)$ implies $R(J_p, (\boldsymbol{x}_i)_{i\in[1,m]}) \le R(\mathrm{GAM}_p(2), (\boldsymbol{x}_i)_{i\in[1,m]})$. Therefore, theorem 26.5 in (Shalev-Shwartz and Ben-David, 2014) implies

$$\mathbb{P}_{(\boldsymbol{x}_i,y_i)_{i\in[1,m]}}\left(\mathbb{E}_{\boldsymbol{x},y}\ell(\hat{f}(\boldsymbol{x}), y) - \ell^* > R(J_p, (\boldsymbol{x})_{i\in[1,m]}) + 5\sqrt{\frac{2\log(2/\delta)}{m}}\right) < \delta.$$

even for J_p, in which $\ell(a, y) = \max\{0, 1 - ay\}$. As $f(\boldsymbol{x})y \in \{\pm 1\}$, it holds that $f(\boldsymbol{x})y > 0 \Leftrightarrow f(\boldsymbol{x})y = 1$ and $f(\boldsymbol{x})y \le 0 \Leftrightarrow f(\boldsymbol{x})y = -1$, which implies $[\![f(\boldsymbol{x})y > 0]\!] = \ell(f(\boldsymbol{x}), y)$ for any $f \in J_p$, $x \in X$, and $y \in Y$. Therefore,

$$\mathbb{E}_{(\boldsymbol{x}_i,y_i)_{i\in[1,m]}}\left(\mathbb{E}_{\boldsymbol{x},y}[\![\hat{f}(\boldsymbol{x})y > 0]\!] - \inf_{f\in J_p}\mathbb{E}_{\boldsymbol{x},y}[\![f(\boldsymbol{x})y > 0]\!]\right) \tag{78}$$

is also an order of $O\left(R(\mathrm{GAM}_p(2), (\boldsymbol{x}_i)_{i\in[1,m]})\right)$ for any distribution of $(\boldsymbol{x}, y) \in X \times Y$.

On the other hand, it is known that, for any $\tilde{f}(\cdot)$ learned from m i.i.d. samples, there exists $\mathbb{P}_{\boldsymbol{x},y}$ such that (78) is an order of $\Omega\left(\sqrt{\mathrm{VCdim}(F)/m}\right)$ under the assumption that $\inf_{f\in F}\mathbb{E}_{\boldsymbol{x},y}[\![f(\boldsymbol{x})y > 0]\!] \neq 0$ (Devroye and Lugosi, 1995; Boucheron et al., 2005). Because J_p contains $2p$ different classifiers, $\mathrm{VCdim}(J_p)$ is at least of $\Omega(\log p)$, which implies that (78) is an order of $\Omega(\sqrt{\log p/m})$.

Therefore, $R(\mathrm{GAM}_p(1), (\boldsymbol{x}_i)_{i\in[1,m]})$ cannot be tighter than $O(\sqrt{\log p/m})$.

Acknowledgments

We thank a bunch of people.

References

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Generalization bounds for learning kernels. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 247–254, 2010.

Luc Devroye and Gábor Lugosi. Lower bounds in pattern recognition and learning. *Pattern recognition*, 28(7):1011–1018, 1995.

Antoine Guisan, Thomas C Edwards, and Trevor Hastie. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological modelling*, 157(2):89–100, 2002.

László Györfi. *Principles of nonparametric learning*. Springer, 2002.

Trevor Hastie and Robert Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in neural information processing systems*, pages 793–800, 2009.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

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

Robert Tibshirani. The lasso method for variable selection in the cox model. In *Statistics in Medicine*, pages 385–395, 1997.

Simon Wood. *Generalized additive models: an introduction with R*. CRC press, 2006.