

Saturating Splines and Feature Selection

Nicholas Boyd

*Department of Statistics
University of California
Berkeley, CA 94720-1776, USA*

NICKBOYD@BERKELEY.EDU

Trevor Hastie

*Department of Statistics
Stanford University
Stanford, CA 94305, USA*

HASTIE@STANFORD.EDU

Stephen Boyd

*Department of Electrical Engineering
Stanford University
Stanford, CA 94305, USA*

BOYD@STANFORD.EDU

Benjamin Recht

*Department of Electrical Engineering and Computer Science
University of California
Berkeley, CA 94720-1776, USA*

BRECHT@BERKELEY.EDU

Michael I. Jordan

*Division of Computer Science and Department of Statistics
University of California
Berkeley, CA 94720-1776, USA*

JORDAN@CS.BERKELEY.EDU

Editor: Jie Peng

Abstract

We extend the adaptive regression spline model by incorporating *saturation*, the natural requirement that a function extend as a constant outside a certain range. We fit saturating splines to data via a convex optimization problem over a space of measures, which we solve using an efficient algorithm based on the conditional gradient method. Unlike many existing approaches, our algorithm solves the original infinite-dimensional (for splines of degree at least two) optimization problem without pre-specified knot locations. We then adapt our algorithm to fit generalized additive models with saturating splines as coordinate functions and show that the saturation requirement allows our model to simultaneously perform feature selection and nonlinear function fitting. Finally, we briefly sketch how the method can be extended to higher order splines and to different requirements on the extension outside the data range.

Keywords: Convex optimization, feature selection, splines, lasso, regression

1. Introduction

Splines—piecewise polynomials with continuity constraints—are widely used to fit data (Hastie et al., 2001, §5.1). One issue with piecewise polynomials is that they behave erratically beyond their boundary knot points, and (typically) grow without bound outside of that

range (Hastie et al., 2001, §5.2). This instability makes extrapolation dangerous; practitioners must take care to avoid querying spline models near or outside of the range of the training data.

Smoothing spline algorithms (De Boor, 2001; Wahba, 1990; Green and Silverman, 1994) ameliorate this problem by fitting *natural splines*, which reduce to a lower-degree polynomial beyond the boundary knots. The most commonly used varieties of smoothing splines are cubic smoothing splines (degree-three splines that reduce to linear outside the boundary knots) and linear smoothing splines, which extend as constant. The saturating splines we propose are closely related to linear smoothing splines.

Smoothing splines use an ℓ_2 or quadratic notion of complexity, and hence fit models with a predetermined and dense set of knot points (Hastie et al., 2001, §5.4). *Adaptive regression splines* (Mammen and van de Geer, 1997), on the other hand, use an ℓ_1 -type penalty, which can result in a sparse set of adaptively chosen knots. However, adaptive regression splines do not reduce to lower degree outside of the range of their largest knots, and hence may suffer from instability.

We propose fitting adaptive regression splines with explicit constraints on the degree of the spline outside of a certain interval. We call such splines *saturating splines*. While the approach we take can be extended to fitting splines of arbitrary degree with constraints on arbitrary derivatives, in this paper we focus on fitting linear splines that are flat (constant) outside the data range; we mention the extension to higher degree splines in §8. We show that saturating splines inherit the knot-selection property of adaptive regression splines, while at the same time behave like natural splines near the boundaries of the data.

We also show a very important benefit of our approach in the context of fitting generalized additive models (Hastie and Tibshirani, 1990) with saturating spline coordinate functions: the saturation constraint naturally results in variable selection. Not only do we control the complexity of each coordinate function through knot selection, but with the saturation condition, no knots on a variable means the variable is out of the model. This is not true for adaptive splines, since the linear term is unpenalized and hence each variable would always be in the model. The lack of feature selection can hurt interpretability and, in certain cases, generalization. The saturation constraint we propose precludes linear functions, and in concert with the adaptive spline ℓ_1 penalty encourages coordinate functions to be identically zero. As a result, generalized additive models fit with saturating spline component functions often depend on only a few input features.

Like smoothing splines and adaptive regression splines, saturating splines arise as solutions to certain natural functional regression problems. We solve the saturating spline fitting problem by reformulating it as a convex optimization problem over a space of measures, roughly speaking, the second derivative of the fitted function. To the best of our knowledge, this approach is novel. We then apply a variant of the classical conditional gradient method (Jaggi, 2013.; Boyd et al., 2017) to this problem. At each iteration of our algorithm, an atomic measure is produced; moreover, we can uniformly bound the number of atoms, which corresponds to the number of knot points in the spline function. (While we manipulate atomic measures, we solve the problem over the space of all measures with finite total variation.) In contrast to standard coordinate descent methods, in each iteration of the conditional gradient method the weights of *two* knot points are adjusted. In the fully corrective step, we solve a finite-dimensional convex optimization problem with ℓ_1

and simple linear constraints. Numerical experiments show that the method is extremely effective in practice.

Our optimization method can exploit warm starts, i.e., it can use an initial guess for the fitted function. This allows us to compute an entire regularization path efficiently, at a cost typically just a small multiple of the effort to solve the problem for one value of the regularization parameter. Because our algorithm is based on the conditional gradient method, we can use the framework of Giesen et al. (2012) to compute a provably ϵ -suboptimal approximate regularization path. When fitting generalized additive models, the regularization path has attractive features: at critical values of the regularization parameter, new regressors are brought into (or, occasionally, out of) the model, or new knot points are added to (or deleted from) one of the existing coordinate functions. Thus our approach combines feature selection and knot point selection.

1.1 Outline

In §2 we introduce a univariate function fitting problem, inspired by the adaptive spline estimation problem of Mammen and van de Geer (1997), that includes the additional requirement that the fitted function saturate. In §3 we make the connection between our function estimation problem and standard adaptive splines, and pose the saturating spline fitting problem as a convex optimization problem over measures. In §4 we modify the classical conditional gradient method to solve this optimization problem. In §5 we extend the optimization problem and algorithm to fit generalized additive models to multivariate data. We illustrate the effectiveness of the method with several examples in §7. We discuss generalizations to higher-degree splines in §8. Finally, we discuss potential extensions and variations in §9. The appendix includes implementation details and proofs.

2. Univariate Function Fitting

We wish to fit a continuous bounded function $f : \mathbf{R} \rightarrow \mathbf{R}$ from data $(x_i, y_i) \in \mathbf{R} \times \mathcal{Y}$, $i = 1, \dots, n$, $x_i \in [0, 1]$. To do this we will choose f to minimize a data mismatch or loss function subject to a constraint that encourages regularity in f , and an additional constraint, saturation, that we describe below.

The loss is given by

$$L(f) = \sum_{i=1}^n \ell(f(x_i), y_i),$$

where $\ell : \mathbf{R} \times \mathcal{Y} \rightarrow \mathbf{R}$ is nonnegative, twice differentiable, and strictly convex in its first argument. Typical loss functions include $\ell(z, w) = (z - w)^2/2$ (standard regression, $\mathcal{Y} = \mathbf{R}$), or $\ell(z, w) = \log(1 + \exp(-zw))$ (logistic regression, with $\mathcal{Y} = \{-1, 1\}$). The loss L is a convex functional of the function f that only depends on the values of f at the data points x_i . The smaller the loss, the better f fits the given data.

We constrain the function f to be simple by limiting the value of a nonnegative regularization functional R . In this paper, we take R to be the total variation of the derivative of f ,

$$R(f) = \text{TV}(f'),$$

a convex functional of f . For a twice-differentiable function f , recall that

$$\text{TV}(f') = \int |f''(x)| dx, \quad (1)$$

i.e., the regularization is the ℓ_1 norm of the second derivative. (As we review in the following section, the modern definition of total variation extends this equality to nondifferentiable functions.) The total variation limit we impose on f is $R(f) \leq \tau$, where τ is a parameter that we use to trade off model fit and model regularity. This regularization constraint implicitly constrains f to be differentiable almost everywhere, with its derivative having finite total variation.

Our model f will be subject to one more constraint, that it saturates (outside the interval $[0, 1]$), which means that it is a (possibly different) constant on the two intervals outside $[0, 1]$: $f(x) = f(0)$ for $x \leq 0$, and $f(x) = f(1)$ for $x \geq 1$. In other words, f extends as a constant outside the nominal data range of $[0, 1]$. In terms of the derivative, this is equivalent to the requirement that f' exists and is zero outside $[0, 1]$.

The fitting problem is then

$$\begin{aligned} & \text{minimize} && L(f) \\ & \text{subject to} && R(f) \leq \tau, \\ & && f'(x) = 0 \text{ for } x \notin [0, 1], \end{aligned} \quad (2)$$

where $\tau \geq 0$ is the regularization parameter. The variable to be determined is the function f , which is in the vector space of continuous functions with derivatives of finite total variation. This fitting problem is an infinite-dimensional convex optimization problem.

In applications the problem (2) is solved for a range of values of τ , which yields the regularization path. The final model is selected using a hold-out set or cross-validation. For $\tau = 0$, f must be constant and the problem (2) reduces to fitting the best constant to the data. As τ increases, f is less constrained, and our fitted model becomes more complex; eventually we expect overfitting. For example, in the case of regression, with a loss function that satisfies $\ell(u, u) = 0$ and data with distinct x_i , the fitting function is the piecewise-linear function that interpolates the data, for large enough τ .

3. Splines and Functions of Bounded Variation

In this section we explore the connection between our fitting problem and degree-one splines, i.e., piecewise-linear continuous functions, which have the form

$$f(x) = c + \sum_{i=1}^K w_i (x - t_i)_+, \quad (3)$$

where $(z)_+ = \max\{z, 0\}$. We assume that the t_i are distinct, and refer to them as knot points or simply knots. The scalars w_i are the weights, and c is the offset. We refer to the function $x \mapsto (x - t_i)_+$ as a hinge function, so a degree-one spline is a finite linear combination of hinge functions, plus a constant.

3.1 Functions of Bounded Variation

A right-continuous function $h : [0, 1] \rightarrow \mathbf{R}$ is of bounded variation if and only if there exists a signed measure μ on $[0, 1]$ with

$$h(z) = \int 1(y \leq z) d\mu(y), \tag{4}$$

where $1(y \leq z) = 1$ for $y \leq z$ and 0 otherwise. The measure μ is unique; we can think of it as the derivative of h . That is, (4) is essentially the second fundamental theorem of calculus with h' replaced by μ .

We also have $\text{TV}(h) = \int d|\mu|$. (This is called the total variation of the measure μ .) We will denote this using the notation $\|\mu\|_1$, to emphasize the similarity with the finite-dimensional case, or the case when h is differentiable: $\text{TV}(h) = \|h'\|_1$. When the measure μ is atomic, the function h is piecewise constant with jumps at the points in the support of μ .

3.2 Splines and Derivatives with Bounded Variation

Now suppose that $f : [0, 1] \rightarrow \mathbf{R}$ has a right-continuous *derivative* of bounded variation. From (4), with $h = f'$, and the fundamental theorem of calculus, we have

$$f(x) = f(0) + \int_0^x f'(z) dz = f(0) + \int_0^x \int 1(y \leq z) d\mu(y) dz \tag{5}$$

$$= f(0) + \int \int_0^x 1(y \leq z) dz d\mu(y) \tag{6}$$

$$= f(0) + \int (x - y)_+ d\mu(y). \tag{7}$$

This shows that any such function is a (possibly infinite) linear combination of hinge functions, plus a constant (i.e., $f(0)$). In this case, the measure μ can be thought of as the *second* derivative of f .

When μ is atomic and supported on a finite set, that is,

$$\mu = \sum_{i=1}^K w_i \delta_{t_i},$$

f is a degree-one spline of the form (3), with $c = f(0)$. So degree-one splines correspond exactly to the case where the measure μ (roughly, the second derivative) has finite support.

We introduce the notation

$$f_\mu(x) = \int_0^x \int 1(t \leq z) d\mu(t) dz = \int (x - t)_+ d\mu(t) \tag{8}$$

to denote the function derived from the measure μ . It is, roughly speaking, the double integral of the measure μ , or the (potentially infinite) linear combination of hinge functions associated with the measure μ . The mapping from μ to f_μ is linear, and we have $\text{TV}(f'_\mu) = \|\mu\|_1$. A simple example of f_μ , its first derivative f'_μ , and its (atomic measure) second derivative μ is shown in Figure 1.

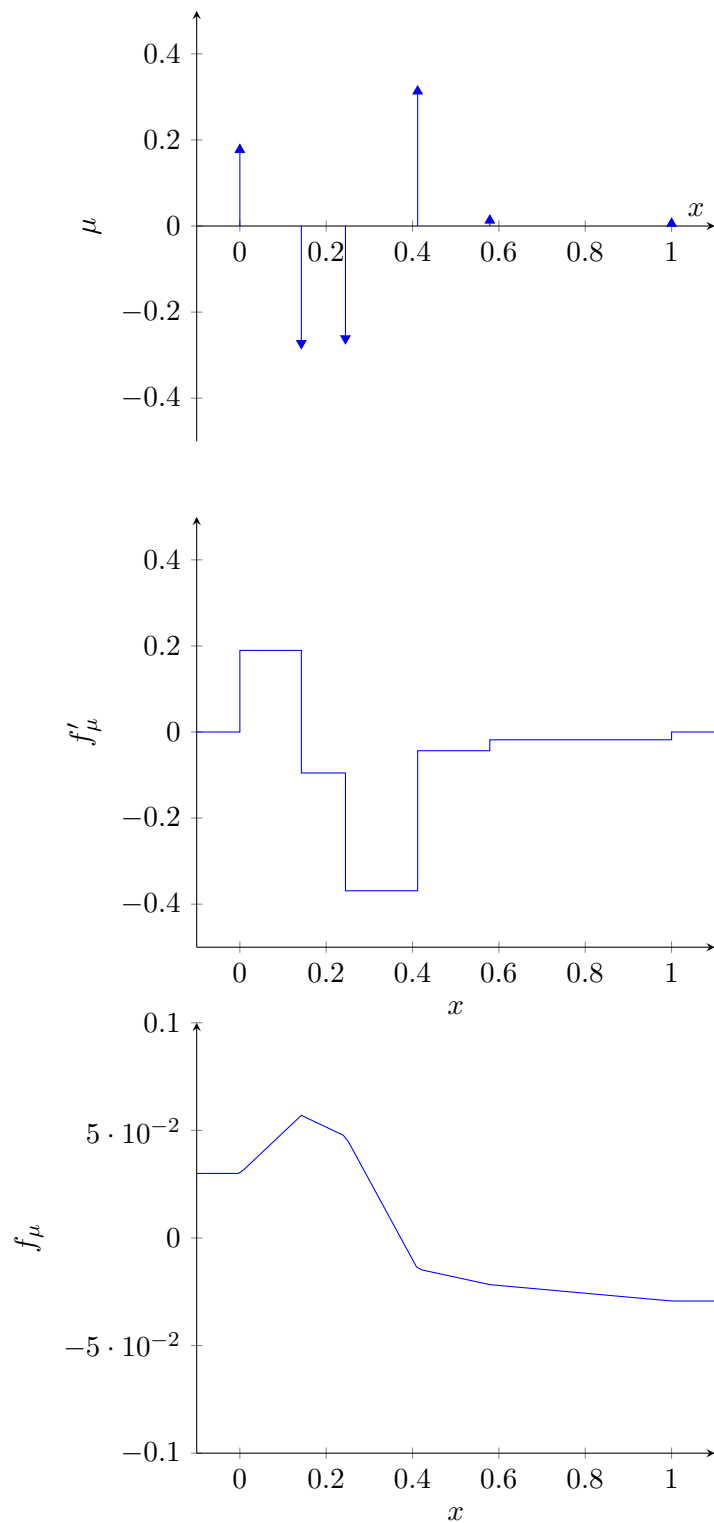


Figure 1: f_μ and f'_μ generated by the atomic measure μ (f''_μ). The regularization functional, $\text{TV}(f'_\mu)$, is the sum of the absolute values of the spikes in μ . Note that the (signed) sum of the spikes in μ is zero: that is, $\int d\mu = 0$, which implies that f_μ saturates.

3.3 Fitting Splines by Optimizing Over Measures

Identifying $f = c + f_\mu$, we can solve the fitting problem (2) by minimizing over the bounded measure μ on $[0, 1]$, and the constant c . The measure μ is the second derivative of f , and the constant c corresponds to $f(0)$. The total variation regularization constraint $\text{TV}(f') \leq \tau$ corresponds to $\|\mu\|_1 \leq \tau$. The saturation condition holds by construction for $x < 0$; to ensure that $f'(x) = 0$ for $x > 1$, we need

$$f'(1) = f'(0) + \int_0^1 d\mu = 0.$$

In other words, saturation of f corresponds to μ having total (net) mass zero. Thus (2) can be rephrased as

$$\begin{aligned} & \text{minimize} && L(E_x \mu + c) \\ & \text{subject to} && \|\mu\|_1 \leq \tau, \\ & && \int d\mu = 0 \end{aligned} \tag{9}$$

over the bounded measure μ on $[0, 1]$, and $c \in \mathbf{R}$. Note the slight abuse of notation here: we now (and for the remainder of the paper) consider L as a functional on \mathbf{R}^n . In the above, E_x is the linear operator that maps μ to the vector $(f_\mu(x_1), \dots, f_\mu(x_n))$, given by (8). E_x is clearly linear, as it is the integral of the function $\psi : \mathbf{R} \rightarrow \mathbf{R}^n$:

$$\psi(t) = ((x_1 - t)_+, \dots, (x_n - t)_+)$$

against μ . We will apply the conditional gradient method directly to this problem.

To gain intuition about the optimization problem (9), we can consider it as a infinite-dimensional analogue of the standard lasso (Tibshirani, 1994). The lasso is the solution to the optimization problem

$$\begin{aligned} & \text{minimize} && \frac{1}{2} \|Aw - y\|_2^2 \\ & \text{subject to} && \|w\|_1 \leq \tau. \end{aligned} \tag{10}$$

Here w is a vector in \mathbf{R}^d , and $A \in \mathbf{R}^{(n,d)}$ is a matrix. Ignoring the constant term c , we see that (9) looks very similar to (10), where E_x plays the role of A ; indeed, E_x is essentially a matrix with n rows and infinitely many columns. Our intuition from the lasso suggests that there should be solutions of (9) that are sparse, which here means that μ is atomic. In terms of f_μ , sparsity means there are solutions of the original functional fitting problem (2) that are degree-one splines. This is indeed the case. Theorem 1 shows that there is a solution of (9) with μ atomic, supported on no more than $n + 2$ points; in other words, f_μ is a degree-one spline with $K \leq n + 2$. Moreover, in practice the solution of (9) will exhibit selection, that is, it will be supported on far fewer than $n + 2$ points.

Theorem 1 *Fix $x_1, \dots, x_n \in [0, 1]$ and $f : \mathbf{R} \rightarrow \mathbf{R}$ with f' (right-continuous) of bounded total variation, and f constant outside of $[0, 1]$. Then there exists a degree-one saturating spline \hat{f} (with at most $n + 2$ knots) that matches f on x_i with $\text{TV}(\hat{f}') \leq \text{TV}(f')$.*

For the remainder of the paper we will ignore the constant term c . It is not difficult to adapt the algorithms we present to handle the constant term, but doing so does add some notational complexity. It's also possible to minimize out c , as it does not affect the regularization term; the resulting problem is still convex in w .

4. The Conditional Gradient Method for Fitting Splines

In this section we outline our algorithm for solving (9) (and therefore also (2)). To that end, we briefly review the classical conditional gradient method (Jaggi, 2013.) and the measure-theoretic version proposed in (Boyd et al., 2017).

The optimization problem we need to solve, (9), (without the constant term c) is

$$\begin{aligned} & \text{minimize} && L(E_x \mu) \\ & \text{subject to} && \int d\mu = 0, \\ & && \|\mu\|_1 \leq \tau. \end{aligned} \tag{11}$$

As noted in the last section, (11) is a convex optimization problem over a space of measures. We closely follow the approach taken in Boyd et al. (2017) and apply the conditional gradient method to this problem directly.

The main benefit of this approach is that we can restrict our attention to *atomic* measures, i.e., μ of the form

$$\mu = \sum_{j=1}^K w_j \delta_{t_j}.$$

Measures of this form are easily representable in a computer, by simply storing a list of (w_j, t_j) pairs. Theorem 1 ensures that the number of knots we need to store is absolutely bounded, i.e., that our algorithm runs in bounded memory. While we manipulate atomic measures, we solve the problem (11) over all bounded measures.

One thing to note about finitely-supported atomic measures is that we can easily optimize over the weights w_j with the knot locations t_j fixed, since this corresponds to a finite-dimensional convex optimization problem amenable to any standard algorithm. Our algorithm makes use of this fact, and alternates between adding pairs of knots and optimizing over the weights w at each iteration. In this latter step knots can be (and indeed eventually must be) removed. In an additional and optional step the knot points can be moved continuously within $[0, 1]$, or to neighboring data points. This step is not needed for theoretical convergence but can improve convergence and the sparsity of the final solution in practice.

4.1 The Conditional Gradient Method

The conditional gradient method (CGM) solves constrained convex optimization problems of the form

$$\begin{aligned} & \text{minimize} && f(x) \\ & \text{subject to} && x \in \mathcal{C}, \end{aligned} \tag{12}$$

with variable $x \in \mathbf{R}^d$. In the above, it is always assumed that the (convex) function f is differentiable. At each iteration of the CGM we form the standard linear approximation to the function f at the current iterate x_m :

$$\hat{f}(x; x_m) = f(x_m) + f'(x - x_m; x_m).$$

Here $f'(d; x)$ is the directional derivative of the function f at x in the direction d , defined by

$$f'(d; x) = \lim_{t \searrow 0} \frac{f(x + td) - f(x)}{t}.$$

Our use of the directional derivative here may seem surprising: for differentiable functions on \mathbf{R}^d , $f'(d; x)$ is always equal to $\langle \nabla f(x), d \rangle$. The direct applicability of directional derivatives to convex functionals of measures motivates us to prefer the directional derivative.

Convexity of f implies that \hat{f} is a *lower* bound on f , that is:

$$\hat{f}(x; x_m) \leq f(x). \tag{13}$$

In the next step of the CGM, we minimize this first-order approximation over the feasible set \mathcal{C} :

$$s_m \in \arg \min_{s \in \mathcal{C}} \hat{f}(s; x_m) = \arg \min_{s \in \mathcal{C}} f'(s; x_m).$$

The point s_m is called the conditional gradient of f . Note that s_m provides a lower bound on $f(x_\star)$:

$$\hat{f}(s_m; x_m) \leq f(x_\star).$$

In particular, we can bound the sub-optimality of the point x_m :

$$f(x_m) - f(x_\star) \leq -f'(s_m - x_m; x_m). \tag{14}$$

One can show (Jaggi, 2013.) that this bound decreases to zero, which means that it can be used as a (non-heuristic) termination criterion. After determining s_m , there are several options for updating x_m . In this paper, we will use the fully-corrective variant of the CGM, which chooses x_{m+1} to minimize f over the convex hull of $\{s_1, s_2, \dots, s_m\}$. Note that this last step may become computationally intensive as k grows, and indeed limits the applicability of the conditional gradient method to problems where this step is computationally feasible. One option is to remove previous conditional gradients as soon as they are not selected in the minimization step. Caratheodory's theorem ensures us that the set of previous conditional gradients we need to track is then bounded by $d + 1$. In practice, however, the algorithm is usually terminated well before $d + 1$ iterations.

Algorithm 1 Fully-corrective conditional gradient method

For $m = 1, \dots$

1. Linearize: $\hat{f}(s; x_m) \leftarrow f(x_m) + f'(s - x_m; x_m)$.
 2. Minimize: $s_m \in \arg \min_{s \in \mathcal{C}} \hat{f}(s; x_m)$.
 3. Update: $x_m \in \arg \min_{x \in \text{conv}(s_1, \dots, s_m)} f(x)$.
-

4.2 Conditional Gradient for Measures

In this subsection, we apply the conditional gradient method to the infinite-dimensional problem (11), which we repeat here:

$$\begin{aligned} & \text{minimize} && L(E_x \mu) \\ & \text{subject to} && \int d\mu = 0, \\ & && \|\mu\|_1 \leq \tau. \end{aligned} \tag{15}$$

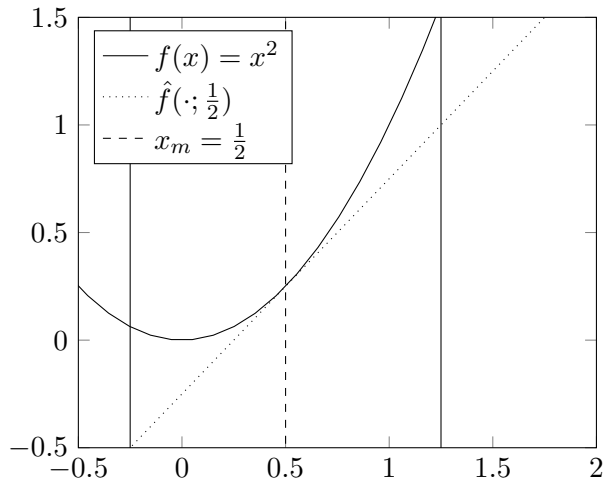


Figure 2: An illustration of a single iteration of the conditional gradient method on the function $f(x) = x^2$ at the point $\frac{1}{2}$. The set \mathcal{C} is the interval $[-0.25, 1.25]$, indicated by the solid vertical lines. The first order approximation $\hat{f}(\cdot; \frac{1}{2})$ is plotted as the dotted line tangential to $f(x)$ at $\frac{1}{2}$. The conditional gradient s_m is the point -0.25 .

First we'll show that the conditional gradient, i.e., the measure s_m , can be chosen to be supported on exactly two points, and is computable in time linear in n . The directional derivative of the objective function in the direction of the measure s at the point μ is given by

$$\begin{aligned} & \lim_{t \searrow 0} \frac{L(E_x(\mu + ts)) - L(E_x\mu)}{t} \\ &= \lim_{t \searrow 0} \frac{L(E_x\mu + tE_x s) - L(E_x\mu)}{t} \\ &= L'(E_x s; E_x\mu) \\ &= \langle \nabla L(E_x\mu), E_x s \rangle_{\mathbf{R}^n}. \end{aligned}$$

We can then interchange the inner-product in $\langle \nabla L(E_x\mu), E_x s \rangle$ with the integral in $E_x s = \int \psi(t) ds(t)$:

$$\langle \nabla L(E_x\mu), E_x s \rangle = \int \langle \nabla L(E_x\mu), \psi(t) \rangle ds(t). \quad (16)$$

Let $g = \nabla L(E_x\mu) \in \mathbf{R}^n$. Note that in the case $\ell(x, y) = \frac{(x-y)^2}{2}$, g is simply the residual $E_x\mu - y$ and $\langle g, \psi(t) \rangle$ is the correlation between the residual and a single hinge function located at t . A conditional gradient is any solution to the following optimization problem

$$\begin{aligned} & \text{minimize} && \int \langle g, \psi(t) \rangle ds(t) \\ & \text{subject to} && \int ds = 0, \\ & && \|s\|_1 \leq \tau. \end{aligned} \quad (17)$$

Without the integral constraint, we would expect there to be a solution to (17) that is a single point-mass: the objective function is the integral of a scalar-valued function against a bounded measure. We'll show that there is always a solution to (17) that is supported on exactly two points. Furthermore, we'll show that those two points can be computed in time linear in n .

First we'll construct a particular feasible point for (17) and then we'll show that it achieves the optimal value. Let

$$t_+ \in \arg \min_t \langle g, \psi(t) \rangle, \quad t_- \in \arg \min_t -\langle g, \psi(t) \rangle.$$

Define

$$s_\star = \frac{\tau}{2} \delta_{t_+} - \frac{\tau}{2} \delta_{t_-}.$$

The objective value achieved by s_\star is

$$o_\star = \frac{\tau}{2} (\langle g, \psi(t_+) \rangle - \langle g, \psi(t_-) \rangle).$$

We'll show that either any measure s that is feasible for (17) has objective value bounded below by o_\star or μ is optimal for (11). Let s be any feasible measure for (17). Decompose s into the difference of two mutually singular non-negative measures: $s = s_+ - s_-$. Then as s is feasible we have $\|s_+\|_1 = \|s_-\|_1 \leq \frac{\tau}{2}$. The objective value achieved by s can be bounded below as follows

$$\begin{aligned} \int \langle g, \psi(t) \rangle ds(t) &= \int \langle g, \psi(t) \rangle ds_+(t) + \int -\langle g, \psi(t) \rangle ds_-(t) \\ &\geq \|s_+\|_1 \left(\min_t \langle g, \psi(t) \rangle \right) + \|s_-\|_1 \left(\min_t -\langle g, \psi(t) \rangle \right) \\ &\geq \|s_+\|_1 \left(\min_t \langle g, \psi(t) \rangle + \min_t -\langle g, \psi(t) \rangle \right). \end{aligned}$$

Suppose $(\min_t \langle g, \psi(t) \rangle + \min_t -\langle g, \psi(t) \rangle) \geq 0$. Then the argument above implies $s_\star = 0$ is a conditional gradient for (11), and thus (14) implies μ is optimal. Otherwise we have

$$\left(\min_t \langle g, \psi(t) \rangle + \min_t -\langle g, \psi(t) \rangle \right) < 0,$$

which implies

$$\|s_+\|_1 \left(\min_t \langle g, \psi(t) \rangle + \min_t -\langle g, \psi(t) \rangle \right) \geq \frac{\tau}{2} \left(\min_t \langle g, \psi(t) \rangle + \min_t -\langle g, \psi(t) \rangle \right) = o_\star.$$

This proves the assertion.

Note that finding t_- and t_+ involves two *separate* optimization problems over $[0, 1]$ instead of one over $[0, 1] \times [0, 1]$. These problems are readily solved by gridding, though in this case they can be solved exactly in time linear in n if we have access to a sorted vector of the data points x_i . To see this, we expand the objective function for t_+ above,

$$t_+ = \arg \min_{0 \leq t \leq 1} \sum_{i=1}^n g_i(x_i - t)_+ = \arg \min_t \sum_{i: x_i \geq t} g_i(x_i - t).$$

If x_i are sorted, we can compute the minimizer between each pair of consecutive data points exactly, since this is simply computing the minimizer of a linear functional over an interval. Thus in a single pass over the data we can compute the global minimizer exactly.

Immediately after computing t_- and t_+ we can use (14) to bound the suboptimality of μ by

$$L(E_x\mu) - L(E_x\mu_\star) \leq - \int \langle g, \psi(t) \rangle d(s_\star - \mu)(t).$$

With this choice of conditional gradient, the fully-corrective step is a finite-dimensional convex problem. Fixing the knot locations encountered as conditional gradients so far, t_1, \dots, t_{2k} , we can do at least as well as the fully-corrective algorithm by solving the following optimization problem:

$$\begin{aligned} & \text{minimize} && L(E_x\mu) \\ & \text{subject to} && \int d\mu = 0, \\ & && \|\mu\|_1 \leq \tau, \\ & && \text{supp}(\mu) \subset \{t_1, \dots, t_{2k}\}. \end{aligned} \tag{18}$$

This is equivalent to the following optimization problem in \mathbf{R}^{2k} :

$$\begin{aligned} & \text{minimize} && L(\sum_j w_j E_x \delta_{t_j}) \\ & \text{subject to} && 1^T w = 0, \\ & && \|w\|_1 \leq \tau. \end{aligned} \tag{19}$$

We can solve this using any of a number of existing algorithms (Boyd et al., 2011; den Berg and Friedlander, 2011). In our implementation we use the conditional gradient method with line-search for simplicity.

By warm starting with an increasing sequence of τ 's, we can efficiently compute an approximate regularization path. Indeed we can even provide a provably ϵ -suboptimal path using the approach of Giesen et al. (2012).

4.3 Convergence

As in the case of ADCG (Boyd et al., 2017) convergence follows immediately from the conditional gradient method proof in general Banach spaces (Dunn and Harshbarger, 1978; Demyanov and Rubinov, 1973; Jaggi, 2013.). The convergence of the conditional gradient method depends on a curvature parameter C_f . C_f is a constant such that the following inequality is satisfied for all $x, s \in \mathcal{S}$ and $\eta \in (0, 1)$:

$$f(x + \eta(s - x)) \leq f(x) + \eta f'(s - x; x) + \frac{C_f}{2} \eta^2.$$

For our purposes $f : \mathbf{R}^n \rightarrow \mathbf{R}$ is simply L and $\mathcal{S} = \{E_x\mu : \|\mu\|_1 \leq \tau, \int d\mu = 0\}$. A simple sufficient condition for C_f to be finite is that ℓ is differentiable with Lipschitz gradient. If C_f is finite, the conditional gradient method converges (in terms of function value) at a rate of at least $1/m$ where m is the iteration counter.

5. Generalized Additive Models

One natural application of univariate splines is fitting generalized additive models (Hastie and Tibshirani, 1990) to multivariate data: $(x_i, y_i) \in \mathbf{R}^D \times \mathcal{Y}$, $i = 1, \dots, n$. That is, fitting a function of the form

$$f(x) = \sum_{d=1}^D f_d(x[d])$$

where each f_d is a simple function from \mathbf{R} to \mathbf{R} (here $x[d]$ is the d -th coordinate of the vector x). We can mimic our approach in the scalar case with the following optimization problem:

$$\begin{aligned} & \text{minimize} && L(f) \\ & \text{subject to} && \sum_d R(f_d) \leq \tau, \\ & && f'_d(x) = 0 \quad \forall x \notin [0, 1], d. \end{aligned} \tag{20}$$

Here R is the same regularizer used in the scalar case, namely

$$R(g) = \text{TV}(g') \simeq \|g''\|_1.$$

As in the scalar case, one can show that there is always an optimal f with each coordinate function f_d a degree-one saturating spline.

This allows us to rephrase (20) as an optimization problem over measures. The only change from the scalar case is that the measure is over the set $\{1, \dots, D\} \times [0, 1]$ —each knot is now attached to a particular coordinate. In other words, we search for a function of the following form:

$$f_\mu(x) = \int (x[d] - t)_+ d\mu(d, t).$$

We again have equality between the ℓ_1 norm of μ and the regularization term:

$$\sum_d R((f_\mu)_d) = \|\mu\|_1.$$

The analogue of (11) is then

$$\begin{aligned} & \text{minimize} && L(E_x \mu) \\ & \text{subject to} && \int \mathbf{1}(d = \hat{d}) d\mu(d, t) = 0, \quad \forall \hat{d} \\ & && \|\mu\|_1 \leq \tau. \end{aligned} \tag{21}$$

The conditional gradient algorithm from the scalar case generalizes immediately to fitting generalized additive models—the only difference is that we now need to find a pair of knots for the same coordinate. This involves solving d pairs of nonconvex optimization problems over $[0, 1]$ —again this can be done by gridding or by sorting the training data.

Saturating splines gain an additional advantage over standard adaptive splines when fitting generalized additive models. The addition of the saturation constraint (that f_d be constant outside of $[0, 1]$) naturally leads to variable selection when fitting generalized additive models. What we mean by variable selection is that the functions f_d are often exactly 0. This is because the saturation constraint means that linear coordinate functions no longer escape the regularization (indeed, they are impossible). This is very different

from the standard adaptive spline setup without the saturation constraint. In that case, linear functions, i.e., $f_d(x[d]) = wx[d]$ completely escape the regularization, and as a result are essentially always included in the model. Linear functions are *not* free with saturation constraints (in fact, outside of the function 0, they are not feasible). When we solve (21) we simultaneously fit nonlinear coordinate functions while doing variable selection.

6. Prior and Related Work

Smoothing splines also have an interpretation as the solution of an infinite-dimensional optimization problem (Hastie et al., 2001, §5.4). In fact, (degree-one) smoothing splines solve

$$\begin{aligned} & \text{minimize} && L(f) \\ & \text{subject to} && \hat{R}(f) \leq \tau, \end{aligned} \tag{22}$$

where

$$\hat{R}(f) = \int f'(x)^2 dx.$$

The solution to (22) is *also* a degree-one natural spline that saturates outside of $[0, 1]$. However, the solutions to (22) and (2) are *very* different. Roughly, (22) is analogous to ridge regression, while (2) is analogous to the lasso. That is, (22) fits functions with as many knots as datapoints, while (2) often fits splines with very few knots.

Another type of spline, that is adaptive but does not saturate, are adaptive regression splines (Mammen and van de Geer, 1997). These splines also arise as solutions to a functional regression problem:

$$\begin{aligned} & \text{minimize} && L(f) \\ & \text{subject to} && \hat{R}(f) \leq \tau, \end{aligned} \tag{23}$$

where

$$\hat{R}(f) = TV(f'(x)).$$

Note that this is (2) without the saturation constraint. Algorithms for solving (23) (for degree-one splines) are based on an extension of Theorem 1, that shows there is a solution to (23) which is actually supported on the data points x_i . Hence a lasso algorithm can be used to find the solution. This also suggests a very simple method to solve our problem (9): we fix the n knot points to be the values of the data x_i , and solve the finite-dimensional convex optimization problem to find the weights. While simple coordinate-descent methods like GLMNet (Friedman et al., 2010) will not immediately work because of the saturation constraint, they could be modified to handle the constraint.

This method does work, but can be much slower than ours since in practice the number of knots is typically much smaller than n for useful values of the regularization parameter τ , and the finite-dimensional problem with n basis functions is very poorly conditioned. With that said, the algorithm we propose—for the piecewise linear case—can be interpreted as a forward active set method for the finite dimensional problem, where we avoid explicitly evaluating all basis functions. One advantage of our measure-theoretic approach is that it immediately generalizes to higher-degree splines, where the support of μ need not be on data points, as we will see in §9. In this case (9) is truly infinite-dimensional, yet our algorithm can still be directly applied.

Trend filtering is a nonparametric function estimation technique, first introduced by Kim et al. (2009), that is very similar to adaptive splines. Indeed, as discussed by Tibshirani (2014), the trend filtering estimate in the constant or piecewise-linear case is exactly the same as the adaptive spline estimate. Trend filtering is increasingly popular as it admits extremely efficient, robust algorithms (Tibshirani, 2014; Ramdas and Tibshirani, 2015). Indeed, some of these algorithms (especially those adapted to fit GAMs, Petersen, Witten, and Simon (2015)) may be adapted to efficiently fit saturating trend filter estimates, which would benefit from the feature selection properties of saturating splines and the computational efficiency of trend filtering.

There are a number of methods for fitting generalized additive models with spline component functions. One approach (Lin and Zhang, 2006) is to use the group-lasso version of (6):

$$R(f) = \sum_d \sqrt{\int f'_d(x)^2 dx}.$$

Extending this idea, Chouldechova and Hastie (2015) use an overlap group-lasso that facilitates selection between zero, linear and nonlinear terms. The differences between these approaches and ours are analogous to the differences between the standard group-lasso and the lasso. While both do feature selection, the penalty functional (6) does not do knot-selection within each coordinate function.

One very similar approach to fitting splines that does not require knot selection (but does not incorporate saturation) is discussed by Rosset et al. (2007).

7. Examples

In all examples we affinely preprocess the data so that all training features lie in $[0, 1]$, and apply the same transformation to the test features (which thus may have values outside of $[0, 1]$). All plots are in terms of the standardized features. For the bone density and abalone data sets we select τ to minimize error on the validation sets. For the Spam and ALS data sets we use cross-validation to estimate τ . We hold out a random subset of size 100 from the training set and train on the remaining data. For each random validation/train split we estimate τ to minimize hold-out error and take our final estimate of τ as the mean over 50 trials.

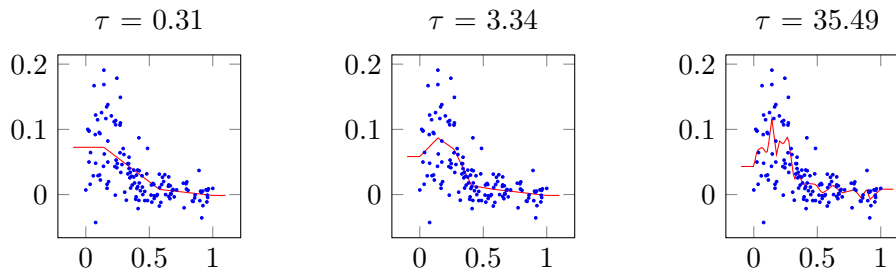


Figure 3: Saturating splines fit to bone density data (shown as scattered points) for 3 values of the regularization parameter τ . *Top*: $\tau = 0.31$; *Middle*: $\tau = 3.34$; *Bottom*: $\tau = 35.45$.

7.1 Bone Density

We start with a simple univariate data set from (Hastie et al., 2001, §5.4). The response variable for this data set is the change in spinal bone density between two doctor visits for female adolescents as a function of age. There are 259 data points, of which we hold out 120 for validation, leaving 139 data points to which we fit a saturating spline. We start with the square loss.

The results are shown in figure 3, for three values of the regularization parameter τ . The scattered points are the training data, the solid line is the saturating spline fit by our algorithm. The figure demonstrates the clear link between τ and the complexity of the optimized spline. Out-of-sample validation suggests setting $\tau \simeq 3.34$, which achieves a validation RMSE of 0.036.

To demonstrate that our proposed method works with more general loss functions, we add 30 simulated outliers to the training set and fit with the pseudo-Huber loss (Charbonnier et al., 1997), a smooth approximation to the Huber loss function given by

$$l_\delta(u) = \delta \left(\sqrt{1 + \frac{u^2}{\delta}} - 1 \right),$$

where $\delta > 0$ is a parameter that interpolates between the absolute value loss and the squared loss. For our experiment we take $\delta = 0.0015$; roughly speaking, the transition between square and linear loss occurs around $\sqrt{\delta} = 0.039$. The results are shown in figure 4. These plots demonstrate that our algorithm can fit losses other than the square loss, and confirms that the pseudo-Huber loss is far more robust to outliers than the basic square loss function. Indeed, on the validation set the least-squares fit achieves a minimum RMSE of 0.096, while the pseudo-Huber fit achieves 0.038, only slightly worse than the fit obtained before the outliers were added to the training data. While this one-dimensional problem is very easy, it shows one advantage of the adaptive spline penalty over smoothing splines: the optimal model has only 5 knot points.

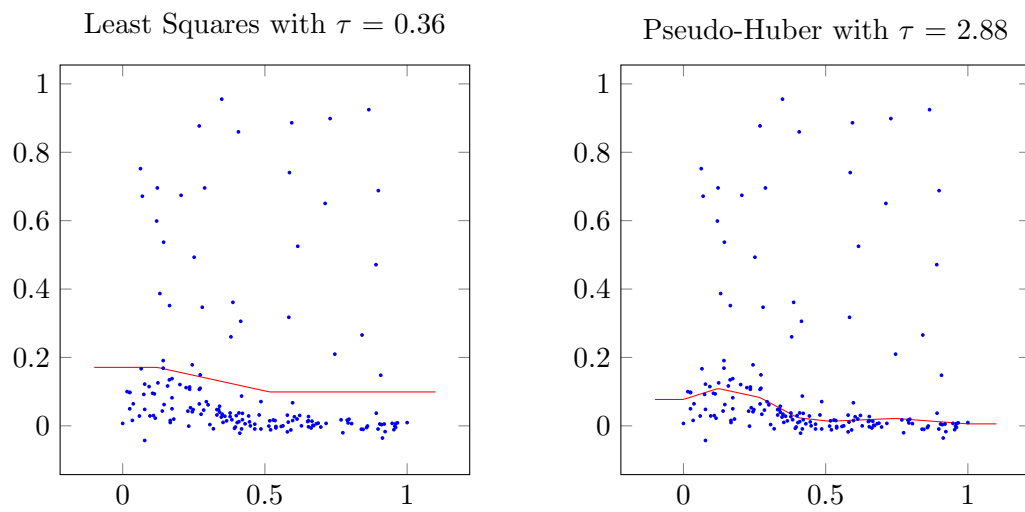


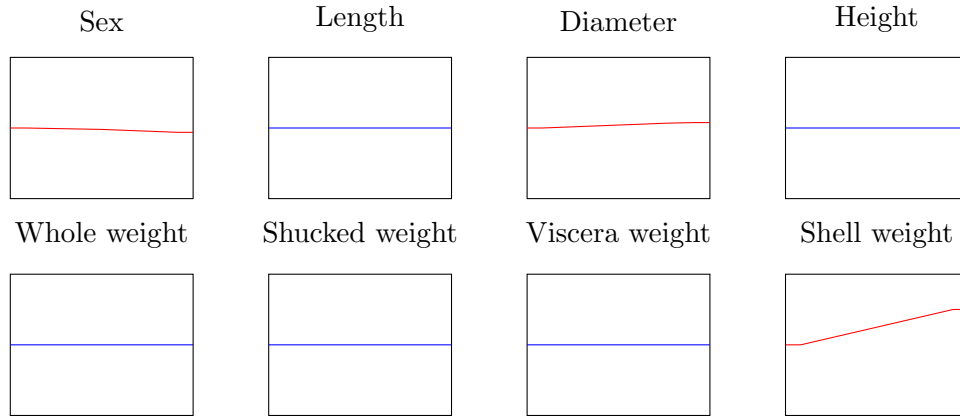
Figure 4: Saturating splines fit to bone density data (shown as scattered points) with simulated outliers for square loss function (left) and pseudo-Huber loss function (right), each for the value of τ that minimizes RMSE on the test set.

7.2 Abalone

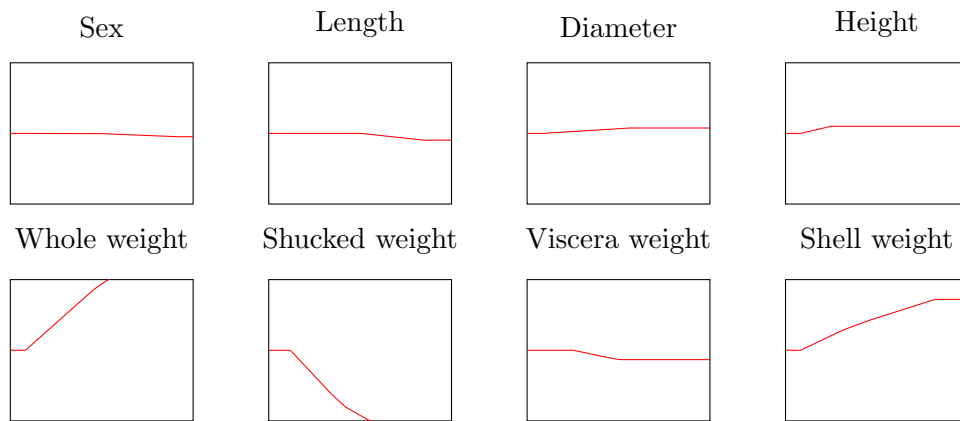
We fit a generalized additive model with saturating spline coordinate functions to the Abalone data set from the UCI Machine Learning Repository (Lichman, 2013). The data consists of 4177 observations of 8 features of abalone along with the target variable, the age of the abalone. We hold out 400 data points as a validation set, leaving 3777 data points to fit the model. The first feature (labeled sex) has three values: Male, Female, and Juvenile, which are coded with values 0, 1, 2; the other 7 are (directly) real numbers. The task is to estimate the age of the abalone from the features.

Cross-validation suggests we choose $\tau \simeq 200$, which achieves a validation set RMSE of 2.131. Because the number of features is low, we can plot the entire generalized additive model. Each plot shows one coordinate function f_d for $d = 1, \dots, 8$ as a function of the standardized feature in $[0, 1]$. The coordinate functions are shown for three values of τ , with the middle one corresponding to the value that minimizes cross-validation RMSE. When a coordinate function is zero, which means that the feature is not used in the model, it is shown in blue. We can see that in the case of strong regularization ($\tau = 20$), several coordinates are not used; for the best model ($\tau = 200$), all features are used, with a few having only a small effect. It is interesting to see how the sex factors into the optimal model. It is neutral on Male or Female, but subtracts a small fixed amount from its age prediction for a Juvenile abalone.

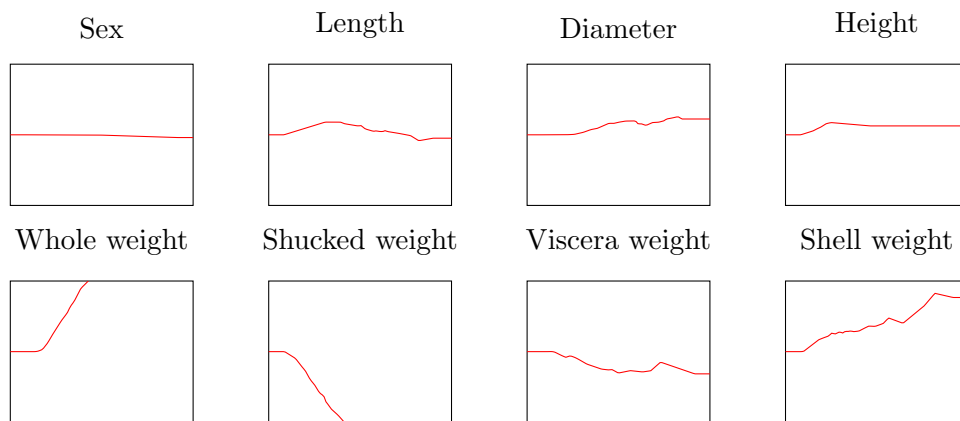
This data set is small enough that we can compare against standard adaptive splines fit using a coarse grid of $[0, 1]$. For this experiment, we fit a GAM with standard adaptive spline component functions using GLMNET (Friedman et al., 2010). The standard adaptive GAM fit, which does no variable selection, achieves a validation set RMSE of 2.137, not significantly worse than the saturating spline model. Our algorithm, however, selects many fewer knot points. The increased number of knots when fitting with GLMNET is perhaps due to the poor conditioning of the gridded problem.



(a) $\tau = 20$



(b) $\tau = 200$



(c) $\tau = 2000$

Figure 5: Coordinate functions for saturating spline generalized additive models fit to Abalone data for three values of the regularization parameter τ .

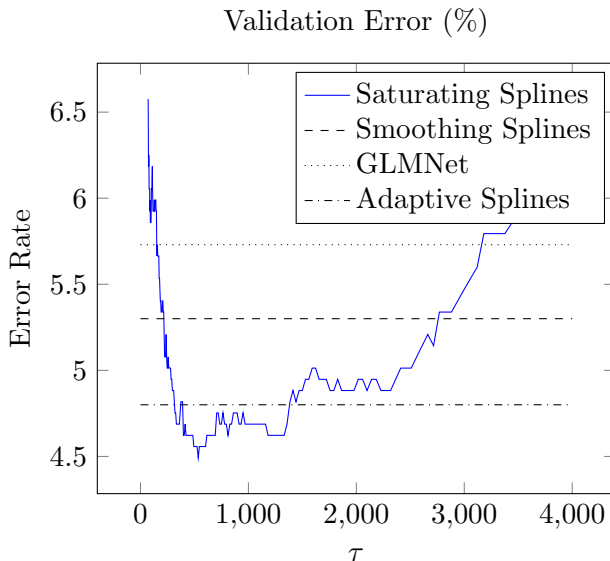


Figure 6: Validation error for saturating spline generalized additive model fit to Spam data set versus regularization parameter τ .

7.3 Spam

We consider the problem of classifying email into spam/not spam, with a data set taken from ESL (Hastie et al., 2001). The data set consists of 57 word-frequency features from 4601 email messages, along with their labels as spam or not spam. Following the approach in ESL (Hastie et al., 2001) we log-transform the features and use the standard train/validation split, with a training set of size 3065, and test set with 1536 samples. We fit a saturating spline generalized additive model with standard logistic loss.

Figure 6 shows the validation error versus the regularization parameter τ . Cross-validation suggests the choice $\tau \simeq 1100$. To show the benefit of nonlinear coordinate functions, we also include the best validation error achieved using a linear model (fit using GLMNet (Friedman, Hastie, and Tibshirani (2010))).

With regularization parameter $\tau = 500$, the model selects 55 of the 57 features. We note that our saturating spline generalized additive model modestly outperforms many methods from ESL (Hastie et al., 2001); for example, smoothing splines yield 5.3% error, while our model has an error rate well below 5%. Figure 7 shows (some of) the coordinate functions for the model with $\tau = 500$. The coordinate functions use very few knots, making them readily interpretable.

For comparison, we fit a GAM with standard adaptive spline coordinate functions. To do so, we grid each dimension with 20 knots and solve the resulting finite-dimensional problem with GLMNET (Friedman et al., 2010). Note that adaptive splines do not penalize linear functions, so there is no feature selection. Adaptive splines achieve a minimum error of 4.8%, significantly worse than saturating splines.

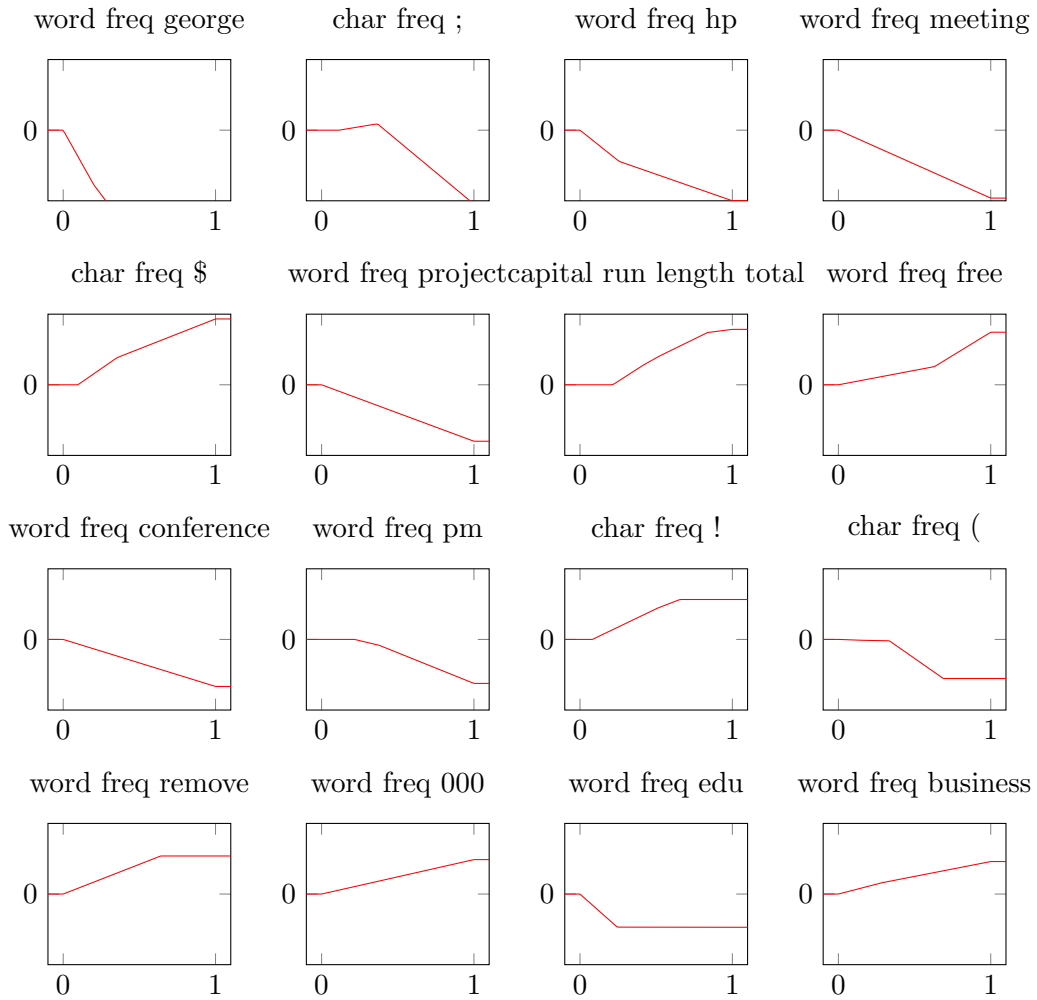


Figure 7: 16 coordinate functions for $\tau = 500$, labeled with the corresponding feature name.

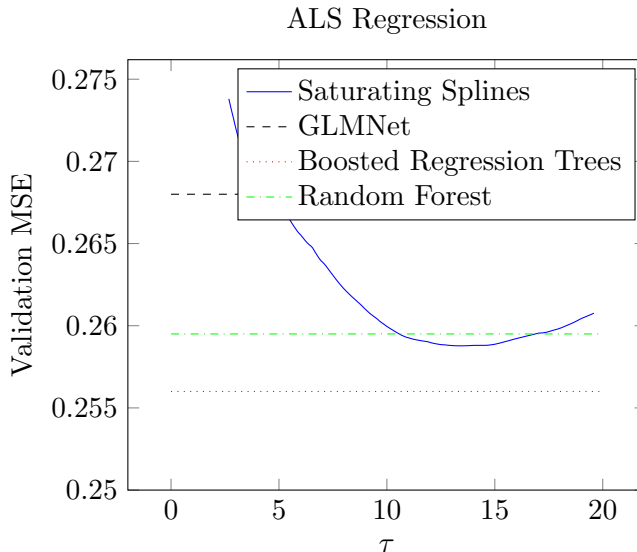


Figure 8: Validation MSE on ALS data set versus regularization parameter τ .

7.4 ALS

Using this data set we try to predict the rate of progression of ALS (amyotrophic lateral sclerosis) in medical patients, as measured by the rate of change in their functional rating score, a measurement of functional impairment. The data set is split into a training set of 1197 examples and a validation set of 625 additional patients. Each datapoint has dimension 369. We fit a generalized additive model with saturating spline component functions to the data using a least-squares objective function. Following (Efron and Hastie, 2016, §17.2), we measure performance using mean-squared error.

We estimate the optimal value of τ using cross validation with a hold-out size of 100 examples and 50 samples; this procedure suggests $\tau = 13$. Figure 8 shows the validation error versus the regularization parameter τ ; the value of τ selected by cross validation achieves low error. On the same plot, we also show the results from Efron and Hastie (2016) using boosted regression trees and random forests. The optimal saturating spline GAM model selects only 50 out of the 369 features, in contrast to boosted regression trees, which use 267. The saturating spline GAM model performs comparably to boosted regression trees and random forests. This is surprising as the saturating spline GAM has no interaction terms. It also uses substantially fewer features, further improving interpretability.

Again we fit a GAM with standard adaptive spline coordinate functions (using GLMNET) to show the advantage of saturation. The standard adaptive spline fit achieves an MSE of 0.547, substantially worse than any other model. We speculate that this is because the unpenalized linear functions lead to immediate overfitting. Indeed, removing the unpenalized linear functions and fitting a model with only hinges gives very similar performance to the saturating spline fit, suggesting that the main advantage of saturation for this application is the removal of the unpenalized linear functions.

Practical advantages of saturating splines These experiments show that saturating splines achieve competitive performance on small classification and regression data sets. In addition, the experiments demonstrate that saturating splines exhibit both knot selection and feature selection—in the context of fitting GAMs. While it is no surprise that saturating splines select fewer knots than smoothing splines (which choose a fully-dense set of knots), it is somewhat surprising that our algorithm selects fewer knots than even *adaptive* splines fit with GLMNET. Finally, the Spam and ALS data sets demonstrate a major advantage of saturating splines over adaptive splines: they simultaneously perform non-linear coordinate function fitting and feature selection. This aids in generalization performance and interpretability. In particular, for the ALS data set saturating spline GAMs achieve *half* the test MSE of adaptive spline GAMs by selecting only 50 of 369 available features.

8. Higher-degree Splines

In the majority of this paper we focused on the functional regression problem (2), with a total variation constraint on the first derivative and a saturation constraint on the zeroth derivative (the function itself). In this section, we consider constraints on higher order derivatives, which lead to solutions that are splines of higher degrees.

$$\begin{aligned} & \text{minimize} && L(f) \\ & \text{subject to} && \text{TV}(f^{(k)}) \leq \tau, \\ & && f^{(k-j)}(x) = 0, \forall x \notin [0, 1]. \end{aligned} \tag{24}$$

We consider the family of nonparametric function estimation problems indexed by $0 \leq j \leq k$. This is the analogue of the functional regression problem (2) with a total variation constraint on the k -th derivative and a saturation constraint on the $(k - j)$ -th derivative. The saturating spline case from the rest the paper is the special case of (24) with $k = 1$, $j = 0$. Widely used cubic natural splines correspond to $k = 3$, $j = 1$. Note that unlike natural splines, which are only defined for some values of j and k , there are no constraints on j and k .

We now show that higher-degree saturating splines solve (24) in general. As $f^{(k)}$ is of bounded TV, there exists a measure μ s.t. $f^{(k)}(x) = \int 1(t \leq x) d\mu(t)$. Then we have

$$\begin{aligned} f^{(k-j)}(x) &= \int \dots \int f^{(k)}(x) dx \dots dx \\ &= \int \dots \int \int 1(t \leq x) d\mu(t) dx \dots dx \\ &= j! \int (x - t)_+^j d\mu(t) + \sum_{l=0}^{j-1} w_l x^l \end{aligned}$$

for some w_l . In the above, all iterated integrals take place j times.

Note that the constraint that $f^{(k-j)}(x) = 0$ for all $x < 0$ implies that the polynomial term, $\sum_{l=0}^{j-1} w_l x^l$ is identically zero. So, we have

$$f^{(k-j)}(x) = j! \int (x - t)_+^j d\mu(t).$$

For $x > 1$, we can remove the nonlinearity, that is, for $x > 1$, $f^{(k-j)}(x)$ is simply the integral of a polynomial in x . We can pull terms involving x out of the integral to get a polynomial in x whose coefficients are nonzero multiples of the first j moments of μ :

$$f^{(k-j)}(x) = j! \sum_{l=0}^j \binom{j}{l} x^{j-k} \int (-t)^k d\mu(t).$$

Again, we note that as this polynomial is identically zero for infinitely many points, all of the coefficients must be zero. In terms of the measure μ , this means:

$$\int t^l d\mu(t) = 0 \quad \text{for } l = 0, \dots, j.$$

This shows that the constraint that the $(k-j)$ -th derivative of f saturate translates to constraints on all moments of μ up to the j -th moment.

While the conditional gradient step becomes more complex with the addition of more moment constraints, the approach taken in this paper can still be applied to (24) as long as j is fairly small—the conditional gradient step for (24) involves a nonconvex optimization problem over $[0, 1]^{j+2}$. This is because we need at least $(j+2)$ point-masses to satisfy the moment constraints. So, fitting quadratic splines that saturate to linear is very easy—in fact the code to do so is essentially identical to that for fitting piecewise linear saturating splines splines—but fitting quadratic splines that saturate to constant is slightly more difficult due to the additional linear constraint on the measure μ . Unfortunately for larger values of j and k , we can no longer hope to find the conditional gradient analytically and must resort to recursive gridding or other global optimization algorithms to find the locations of the new knots.

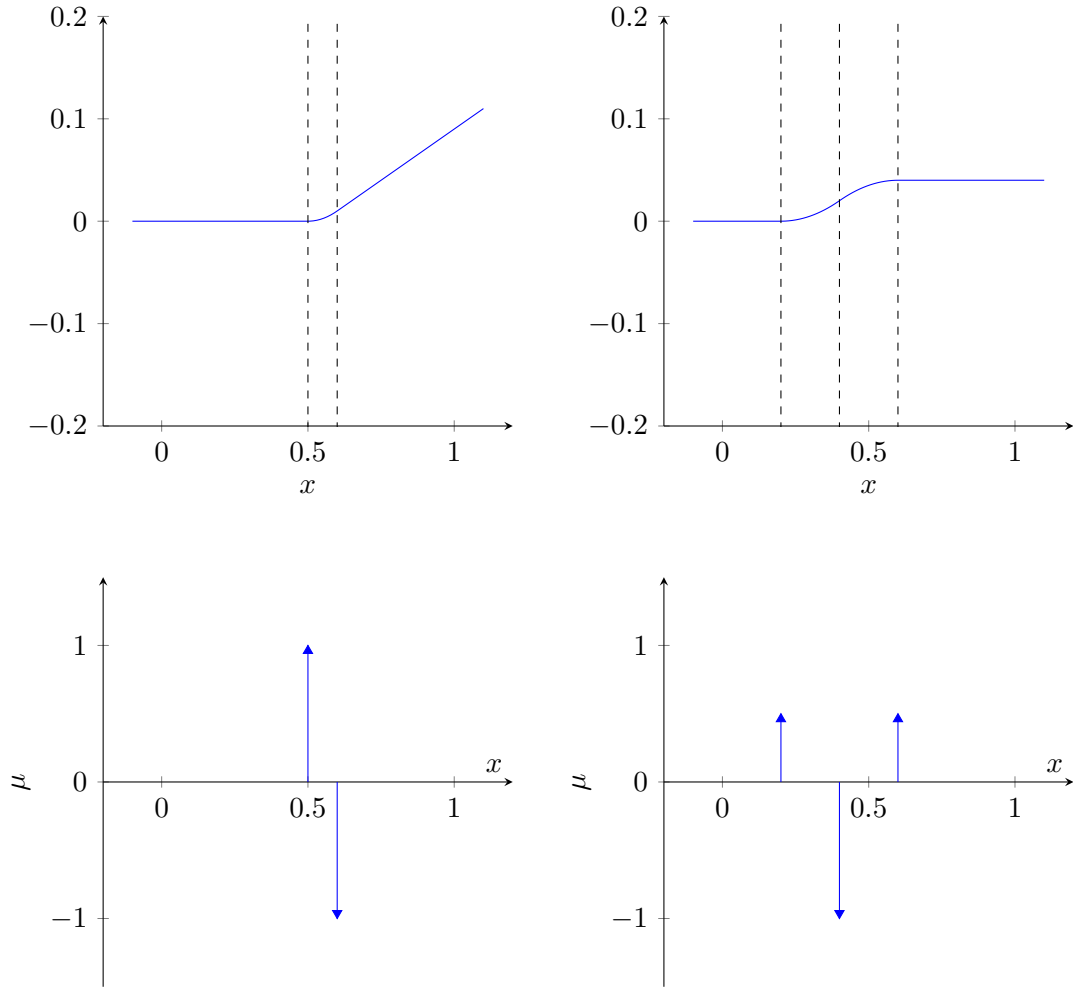


Figure 9: The top two plots show conditional gradients for $k = 2$ with $j = 0$ and $j = 1$ respectively. The dashed lines denote the locations of the point masses: when $j = 1$, the conditional gradient consists of three point masses. The bottom plots show the corresponding measures.

9. Variations and Extensions

While saturation is often a natural prior, the approach we take in this paper can also be applied to other (convex) variations on (9). For example, we could add the constraint that the fitted function is monotone nondecreasing, or takes values in a given interval.

A simple algorithmic extension would be to incorporate nonconvex optimization in the spirit of Boyd et al. (2017). At each iteration we adjust the weights of the atomic measure (w), but we could also adjust the knot locations (t). The objective in (19) is nonconvex in t_i , but we can still attempt to find a local minimum. As long as we do not increase the objective function the algorithm is still guaranteed to converge (Boyd et al., 2017). In the case of degree one splines, we can use the fact that the knot points can, without loss of generality, be chosen to be on the data points to make discrete adjustments to the knot locations.

To fit vector-valued functions, for example in multiclass classification, we would need to extend (9) to use *vector-valued* measures. This is the natural measure-theoretic analogue to the group-lasso.

In multivariate fitting problems with significant interactions between features generalized additive models may underfit. One possible solution is to use single-layer neural networks: i.e., learn functions of the form

$$x \mapsto \sum_{i=1}^K w_i (v_i^T x - t_i)_+.$$

In the above, v_i are constrained to lie in the unit ball. Unfortunately, the conditional gradient step for networks of this form is NP-hard (Bach, 2017). In many practical applications, however, we might expect that the degree of the interaction is bounded. That is, each v_i has bounded cardinality. If we assume $\|v_i\|_0 \leq 2$, i.e., we only fit pairwise interactions, we can still apply the conditional gradient method. In this case, the fitting function is a sum of functions of pairs of the variables, formed from the basis elements

$$((\cos \theta)x_p + (\sin \theta)x_q - t)_+,$$

with (continuous) parameters θ and t and (index) parameters p and q (i.e., $v = (\cos \theta)e_p + (\sin \theta)e_q$). (This is practical only if d is small enough.) Such functions capture nonlinear relationships between (pairs of) variables.

10. Conclusion

In this paper we propose a modification of the adaptive spline regression model—namely saturation constraints. We show that saturating splines inherit knot-selection from adaptive splines, and have a very important quality in the context of generalized additive models: feature selection. This allows saturating spline generalized additive models to remain interpretable and (crucially) avoid overfitting when applied to multivariate data. We also propose a simple, effective algorithm based on the standard conditional gradient method for solving the saturating spline estimation problem with arbitrary convex losses. Finally, we apply our algorithm to several data sets, demonstrating the simplicity of the resulting models.

Acknowledgments

We would like to thank Aaditya Ramdas for many helpful discussions about trend filtering. NB was generously supported by a Google Fellowship from the Hertz Foundation.

Appendix A. Implementation Details

We provide a simple, unoptimized implementation in the Rust language. The runtime of our algorithm is dominated by the fully-corrective step, that is, solving the finite-dimensional convex optimization problem (19). We solve (19) using a proximal Newton method and the standard conditional gradient method with exact linesearch. To be precise, at each iteration, we form the second-order approximation to the objective function

$$f(w) \simeq C + (w - \hat{w})^T \nabla_w f(\hat{w}) + \frac{1}{2} (w - \hat{w})^T \nabla^2 f(\hat{w}) (w - \hat{w})$$

which we then minimize (over the constraint set) using the standard conditional gradient method with (exact) linesearch. Note that this is a Newton step with fixed step-length of 1: as in GLMNET (Friedman et al., 2010), we omit a line search in the interest of speed.

We chose to use a proximal Newton method because of its relative simplicity; other standard convex optimization algorithms may give much better practical performance, especially when the number of data points, n , is extremely large.

Appendix B. Saturating Hinges

In this section we introduce a heuristic for solving an approximation to (9) using existing algorithms for the lasso. Here we consider the case where $\hat{R}(f) = TV(f'(x))$, and hence as pointed out in Section 6 the solution is an expansion in piecewise linear splines with knots at the unique data points. Let h_j be a hinge function at knot t_j : $h_j(x) = (x - t_j)_+$, and suppose we have knots $t_1 < t_2 < \dots < t_k$. Define $f(x) = w_0 + \sum_{j=1}^k w_j h_j(x)$. Given a sample $T = \{(x_i, y_i)\}_1^N$, solving (9) amounts to solving

$$\underset{w_0, w}{\text{minimize}} \left\{ \sum_{i=1}^N \ell(y_i, f(x_i)) + \lambda \|w\|_1 \right\} \quad \text{s.t. } f'(t_k) = 0. \quad (25)$$

Here we've exchanged a constraint on the total variation of $TV(f'(x))$ with a penalty. The condition $f'(t_k) = 0$ is equivalent to $\sum_{j=1}^k w_j = 0$. If the points x_i are unique, then $k = n$; irrespective $t_1 = \min(x_i)$, $t_k = \max(x_i)$ and by construction the estimate \hat{f} is constant beyond the data.

Without the gradient condition $f'(t_k) = 0$, solving (25) amounts to a large lasso problem, for which efficient software is available. Here our goal is to transform the problem to get rid of this constraint. For more generality we do this for an arbitrary set of ordered knots.

Suppose $k < n$ and the right-most knot is inside the range of the data. Consider the following formulation. Let $s_j(x) = h_j(x) - h_k(x)$, a ‘‘saturating’’ hinge function. It looks like a piecewise linear sigmoid, and goes horizontal at t_k (see figure 10).

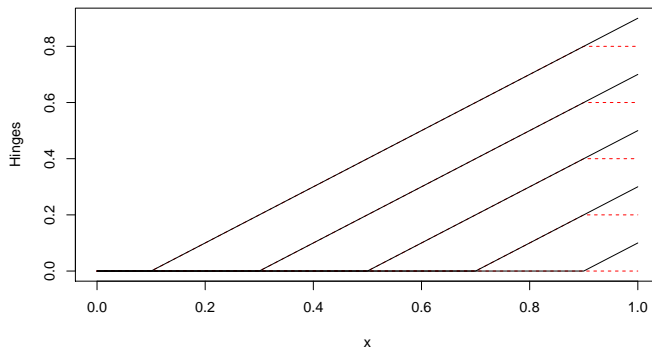


Figure 10: Hinges and saturating hinges.

Without the ℓ_1 constraint (or when $\lambda = 0$) the solution to (25) is equivalent to the solution to the problem with the reduced basis $g(x) = \theta_0 + \sum_{j=1}^{k-1} \theta_j s_j(x)$:

$$\underset{\theta_0, \theta}{\text{minimize}} \sum_{i=1}^N \ell(y_i, g(x_i)). \quad (26)$$

This is easy to see. f is an affine expansion in the h_j , and hence any nonsingular $k \times k$ transformation C of the vector of functions $h(x) = (h_1(x), h_2(x), \dots, h_k(x))$ spans the same space. It is easy to see that with $s(x) = C^T h(x)$, and

$$C = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ -1 & -1 & \cdots & -1 & -1 \end{bmatrix}$$

that $s_j(x)$ are as described (and $s_k(x) = -h_k(x)$). Now $h(x)^T w = h(x)^T C C^{-1} w = s(x)^T \theta$, with $\theta = C^{-1} w$. However, in this case $C^{-1} = C$, and hence $\theta_j = w_j$, $j = 1, \dots, k-1$ and $\theta_k = \sum_{j=1}^k w_j$. In this new basis, imposing the constraint amounts to setting $\theta_k = 0$, or simply deleting the last basis function. So fitting the linear model f subject to $f'(t_k) = 0$ is equivalent to fitting the model g without constraints, and in fact the $\theta_j = w_j$, $j = 1, \dots, k-1$.

So in summary, fitting a constrained optimization with the hinge functions is equivalent to fitting an unconstrained optimization with the reduced set of saturating hinges. The remaining question is does this also work with the penalty $\lambda \|w\|$ as in (25). Not quite, but close. It turns out we are still missing a penalty term $\lambda |\sum_{j=1}^{k-1} \theta_j|$. Hence the transformed problem is

$$\underset{\theta_0, \theta \in \mathbf{R}^{k-1}}{\text{minimize}} \left\{ \sum_{i=1}^N \ell(y_i, g(x_i)) + \lambda \|\theta\| + \lambda \left| \sum_{j=1}^{k-1} \theta_j \right| \right\} \quad (27)$$

If we are willing to ignore this last penalty, we can fit the saturated spline model using a fast lasso solver, such as `glmnet`. By generating such a basis for each variable in a GAM, this same approach can be used to fit a saturated GAM regularization path.

The impact is that one could fit a saturated gam model by running say `glmnet` on the $\{s\}$ bases. An example is given on the spam data in figure 11, where the regularization path was computed at a 100 values of λ in seconds. There are of course some caveats.

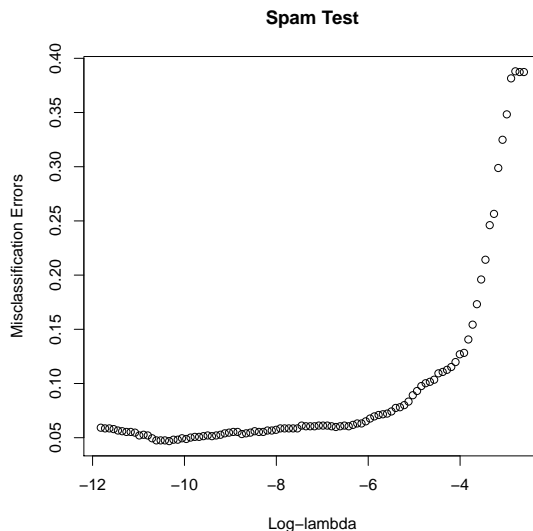


Figure 11: Performance on the spam test data, with 20 knots per variable. Increasing to 50 did not make much difference. Minimum error is 0.047.

- If you use all the knots for each of p variables, you end up with a data matrix of dimension $N \times Np$, which does not scale too well. So instead one might use a smaller grid of knots; e.g., map the variables onto $[0, 1]$, and then use a grid of say 50 evenly spaced knots on this grid, including the end knots.
- With a large number of knots, the “variables” are highly correlated, and this can cause numerical issues. The main issue we see is that active sets tend to be larger than they should be.

Nevertheless, this is an alternative algorithm, which is closer in spirit to the adaptive splines algorithm.

Appendix C. Proof of Theorem 1

Theorem 1 Fix $x_1, \dots, x_n \in [0, 1]$ and $f : \mathbf{R} \rightarrow \mathbf{R}$ with f' (right-continuous) of bounded total variation, and f constant outside of $[0, 1]$. Then there exists a degree-one saturating spline \hat{f} that matches f on x_i with $\text{TV}(\hat{f}') \leq \text{TV}(f')$.

Proof Without loss of generality, we will assume $f(0) = 0$. Let $\tau = \text{TV}(f')$. As f' has bounded total variation, there exists a measure μ on $[0, 1]$ such that $f(x) = f_\mu$:

$$f(x) = \int (x - t)_+ d\mu(t).$$

That is, f is a spline with infinitely many knots. The idea is to use Caratheodory's theorem for convex hulls to see that, as we only care about μ in terms of its action on a finite number of functions (basically, we only care about the values of f at x_i), we can replace μ with a measure supported on finitely many points.

To make this idea rigorous, note that the vector

$$v = (f(x_1), \dots, f(x_n), 0) = \int ((x_1 - t)_+, \dots, (x_n - t)_+, 1) d\mu(t)$$

must lie in convex hull of the (convex) set

$$C = \{\pm(\tau(x_1 - t)_+, \dots, \tau(x_n - t)_+, \tau) : t \in [0, 1]\} \subset \mathbf{R}^{n+1}$$

as $\|\mu\|_1 = \tau$. Caratheodory's theorem for convex hulls ensures us that v can be represented as a convex combination of at most $n + 2$ points from C . Letting these $n + 2$ points be represented by their indicies, t_1, \dots, t_{n+2} , and their weights $\alpha_1, \dots, \alpha_{n+2}$ we define $w_j = \alpha_j \tau$ to obtain:

$$\begin{aligned} f(x_i) &= \sum_j w_j (x_i - t_j)_+ = f_\mu(x_i) \\ \sum_j w_j &= 0. \end{aligned}$$

Here $\mu = \sum_j w_j \delta_{t_j}$. As $\sum_j |w_j| = \tau$, we have $\text{TV}(f'_\mu) = \|\mu\|_1 = \tau$. ■

References

- F. Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.
- N. Boyd, G. Schiebinger, and B. Recht. The alternating descent conditional gradient method for sparse inverse problems. *SIAM Journal on Optimization*, 27(2):616–639, 2017.
- S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.
- P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *Trans. Img. Proc.*, 6(2):298–311, February 1997. ISSN 1057-7149.

- A. Chouldechova and T. Hastie. Generalized additive model selection. *ArXiv e-prints*, June 2015.
- C. De Boor. *A Practical Guide to Splines*. Applied mathematical sciences. Springer, Berlin, 2001.
- V. F. Demyanov and A. M. Rubinov. Approximate methods in optimization problems. *Journal of Applied Mathematics and Mechanics*, 53(7):499–499, 1973. ISSN 1521-4001.
- E. Van den Berg and M. P. Friedlander. Sparse optimization with least-squares constraints. *SIAM Journal on Optimization*, 21(4):1201–1229, 2011.
- J.C. Dunn and S. Harshbarger. Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62(2):432 – 444, 1978. ISSN 0022-247X.
- B. Efron and T. Hastie. *Computer Age Statistical Inference*. Institute of Mathematical Statistics Monographs. Cambridge University Press, 2016. ISBN 9781107149892.
- J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. ISSN 1548-7660.
- J. Giesen, M. Jaggi, and S. Laue. Approximating parameterized convex optimization problems. *ACM Trans. Algorithms*, 9(1):10:1–10:17, December 2012. ISSN 1549-6325.
- P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. Monographs on statistics and applied probability. Chapman & Hall, Boca Raton, London, New York, 1994. ISBN 0-412-30040-0.
- T. Hastie and R. Tibshirani. *Generalized Additive Models*, volume 43. CRC Press, 1990.
- T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.
- M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. *ICML*, 2013.
- S. Kim, K. Koh, S. Boyd, and D. Gorinevsky. ℓ_1 trend filtering. *SIAM review*, 51(2): 339–360, 2009.
- M. Lichman. UCI machine learning repository, 2013. URL <http://archive.ics.uci.edu/ml>.
- Y. Lin and H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.*, 34(5):2272–2297, 10 2006.
- E. Mammen and S. van de Geer. Locally adaptive regression splines. *Ann. Statist.*, 25(1): 387–413, 02 1997.
- A. Petersen, D. Witten, and N. Simon. Fused lasso additive model. *Journal of Computational and Graphical Statistics*, pages 1–37, 2015.

- A. Ramdas and R.J. Tibshirani. Fast and flexible admm algorithms for trend filtering. *Journal of Computational and Graphical Statistics*, 2015.
- S. Rosset, G. Swirszcz, N. Srebro, and J. Zhu. ℓ_1 regularization in infinite dimensional feature spaces. pages 544–558, 2007.
- R.J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.
- G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.