Statistica Sinica **31** (2021), 473-489
doi:https://doi.org/10.5705/ss.202018.0144

CONVERGENCE RATES OF NONPARAMETRIC PENALIZED REGRESSION UNDER MISSPECIFIED SMOOTHNESS

Noah Simon and Ali Shojaie

University of Washington

Abstract: We present a general approach for computing the convergence rates of nonparametric penalized regression estimators under misspecified smoothness, where the true regression function lies in the closure, but not the interior, of the space of smooth functions characterized by the penalty. The proposed approach uses an approximating/representative sequence that has a finite (but growing) penalty value. Here, to establish consistency, we balance the rate at which the penalty grows, with the approximation error of the representative sequence. We apply these ideas to the two most commonly used nonparametric penalties: total variation and Sobolev semi-norms. We give an upper bound for the rate at which we can estimate a function that exhibits bounded $l+1$th-order total-variation or Sobolev complexity, using a kth-order total-variation or Sobolev penalty (for $k > l+1$) respectively. Our bounds have a simple form that depends on k and l. In particular, we show that using total-variation penalties, we will achieve a rate better than $n^{-1/2}$ for any $l \geq 0$ and $k \geq 1$. We evaluate the sharpness of our bounds based on total-variation penalties using a simulation. Empirically, for $l = 0$ our bound appears to be sharp; however for $l \geq 1$, there appears to be a small gap between our upper bound and the empirical rate.

Key words and phrases: Misspecification, non-parametric estimation, penalized regression, sobolev, total-variation.

1. Introduction

Suppose we observe independently drawn (x, y) pairs, with $x \in \mathcal{X}$, $y \in \mathbb{R}$, and

$$y_i = f^*(x_i) + \epsilon_i, \tag{1.1}$$

where ϵ_i are independent, with $\mathrm{E}[\epsilon_i | x_i] = 0$, and $\mathrm{var}(\epsilon_i | x_i) = \sigma_\epsilon^2$ for all i. Further suppose, we aim to estimate the regression function f^*. Rather than assuming that f^* has a specific parametric form, we can estimate it more flexibly by making

Corresponding author: Noah Simon, 1705 Northeast Pacific Street University of Washington, Box 357232 Seattle, WA 98195, USA. E-mail: nrsimon@uw.edu.

an assumption about its smoothness or structure. For instance, we can assume that f^* has k bounded derivatives, or that it is monotone. These assumptions can be encoded using a functional, $P(f)$, which is small (or finite) when f has the desired structure, and large (or infinite) when it does not. Penalized regression is a popular method for estimating f^* in this context (van de Geer (2000)). The penalized estimator is given by

$$\hat{f} \leftarrow \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{2} \|y - f\|_n^2 + \lambda_n P(f), \tag{1.2}$$

where $\|y - f\|_n^2 \equiv (1/n) \sum_{i=1}^n (y_i - f(x_i))^2$ denotes the empirical norm, $\lambda_n > 0$ is a penalty parameter, and $\mathcal{F}$ is a class on which $P(\cdot)$ is defined. In many cases, the penalized estimate $\hat{f}$ has good properties. In particular, if $f^* \in \mathcal{F}_P^C \equiv \{f : \mathcal{X} \to \mathbb{R} \text{ s.t. } P(f) \leq C\}$, for some C, and $\mathcal{F}_P^C$ is not too large, then, for properly chosen λ_n, $\hat{f}$ is a rate-optimal estimator for f^* (in a minimax sense over $\mathcal{F}_P^C$ (van de Geer (2000))).

When $\mathcal{X} = \mathbb{R}$, a popular classical choice is the kth-order Sobolev penalty, $\int_x \left|f^{(k)}(x)\right|^2 dx$. Using this penalty, (1.2), can be calculated by solving a simple linear system (Craven and Wahba (1978)). We consider the following seminorm version of the Sobolev penalty: $P_k^2(f) \equiv [\int_x \left|f^{(k)}(x)\right|^2 dx]^{1/2}$. There is an equivalence between the solutions to (1.2) with $P_k^2(f)$, and $[P_k^2(f)]^2$, though the appropriate choice of λ_n is different for the two penalties (van de Geer (2000)).

An increasingly popular modern choice is the kth-order total-variation seminorm: $P_k(f) \equiv \sup_{x_1,\ldots,x_M} \sum |f^{(k-1)}(x_{m+1}) - f^{(k-1)}(x_m)|$, where the supremum is taken over all partitions $x_1 < \cdots < x_M$. If the function f has a kth-order derivative, then $P_k(f) = \int_x \left|f^{(k)}(x)\right| dx$. This penalty is popular because (a) for f^*, a piecewise $(k-1)$-order polynomial, it permits a faster convergence rate than that of a Sobolev penalty, (b) it results in fitted functions that are $(k-1)$-order splines with data-adaptively determined knots (Tibshirani (2014)), and (c) the fitted function can be obtained from asymptotically equivalent formulations that are efficient to compute (Mammen and van de Geer (1997); Ramdas and Tibshirani (2016)).

If $f^* \in \mathcal{F}_k \equiv \{f : \mathbb{R} \to \mathbb{R} \,|\, f \text{ is } k \text{ times weakly differentiable}, P_k(f) < \infty\}$, it is well known that solving (1.2) with $P = P_k$ results in an estimate $\hat{f}$ that achieves the minimax rate, $\int_{\mathcal{X}} (\hat{f}(x) - f^*(x))^2 dP(x) = O_p(n^{-2k/(2k+1)})$, for functions over this class with $P_k(f) \leq P_k(f^*)$ (Tibshirani (2014)). In practice, one will rarely, if ever, know that $P_k(f^*) < \infty$. Nonetheless, the penalized estimator (1.2) is still commonly used, often assuming a certain order k of differentiability. It is thus

natural to ask whether $\hat{f}$ is still a sensible estimator when these differentiability conditions are not met.

In this study, we explore what happens if we use penalty P_k when, in fact, $f^* \notin \mathcal{F}_k$. In particular, we examine the behavior of the estimator when $f^* \in \mathcal{F}_{l+1}$ for some $l+1 < k$. We find that the penalized estimator (1.2) remains consistent, and give an upper bound on its convergence rate. Our upper bound for the rate depends on both k and l; however, for all k and l, the bound obeys $\|\hat{f} - f\|_n^2 = o_p\left(n^{-1/2}\right)$. To prove these results, we discuss a simple and general framework for bounding the mean squared error (MSE) of penalized estimators when f^* lies in the closure of $\mathcal{F}_P^\infty = \{f \mid P(f) < \infty\}$. Next, we use this general framework to derive results for functions of bounded variation. Then, we extend these results to upper bound convergence rates in two additional scenarios: 1) the true regression function lies in the lth-order Sobolev class (i.e., $P_l^2(f^*) < \infty$), but we do not have $f^* \in \mathcal{F}_k$, and the total-variation penalty P_k is used; and 2) a higher-order Sobolev penalty is used in the estimation, but the true f^* lies in a lower-order Sobolev or total-variation class. The theoretical tools we use are not new (Mammen and van de Geer (1997); van de Geer (2000)): However, to the best of our knowledge, their application to rates of convergence when there is a mismatch between the penalty-induced structure and the true structure is novel. This can be applied to many nonparametric regression problems: For most P that encode smoothness, the ℓ_2-closure of $\mathcal{F}_P^\infty$ contains all square integrable functions.

Somewhat similar ideas in approximation theory have been used previously in the context of nonparametric estimation. In particular, wavelet approximations have been developed for estimation in Besov spaces (Donoho and Johnstone (1995)), and ridgelets and curvelets have been proposed for more general multivariate spaces (Candes (1998); Starck, Candès and Donoho (2002)). In addition, approximation results have been given for neural networks (Barron (1993)). To the best of our knowledge, however, these ideas have not previously been applied to nonparametric estimation based on penalized regression, where the structure induced by the penalty does not match the true structure of the underlying regression function.

For low dimensional problems, there are so-called *adaptive estimation procedures* that achieve near minimax rates over a collection of orders (e.g., kernel smoothers based on Lepski procedures (Lepski and Spokoiny (1997)), and some wavelet-thresholding estimators (Donoho and Johnstone (1995))). Nevertheless, our results are useful. In practice, penalized regressions with Sobolev or total-

variation-based penalties are still common (despite being nonadaptive). Thus, it is of interest to analyze the performance of procedures employed in the current practice of statistics (Szczesniak et al. (2013); Saâda-Bouzid et al. (2017); Omranian et al. (2016); Tibshirani (2014)). Furthermore, many adaptive procedures (e.g., Lepski-type) are difficult to employ in higher-dimensional additive/sparse-additive models, where the degree of smoothness may vary by component. Although we do not analyze such scenarios directly, the ideas presented in this paper may be extended to do so.

2. Framework

Suppose we have n pairs of data (x_i, y_i) generated as in (1.1), although with $\mathcal{X} = [-1, 1]$. Furthermore suppose $\epsilon_i \sim N(0, \sigma^2)$ are drawn independently (although our results require only subGaussian tails). Here, we estimate f^* using $\hat{f}$ by solving the penalized regression problem given in (1.2), for a given choice of P and λ_n. Let $\mathcal{F} = \{f : \mathbb{R} \to \mathbb{R} \mid P(f) < \infty\}$.

Before moving on, we give an intuitive overview of our approach. Our approach has two components:

i) Suppose $f^* \notin \mathcal{F}$. For an arbitrarily chosen candidate *representative* $f^O \in \mathcal{F}$, we show that the error from a penalized regression comes from two sources: (a) the distance from f^O to f^*, and (b) the rate at which we could have estimated f^O if it were the true conditional mean. Here, the choice of penalty P is captured in the latter source.

ii) We choose a sequence of *representatives* $f_0^O, f_1^O, \ldots$ in $\mathcal{F}$ that converges to f^*. Because $f^* \notin \mathcal{F}$, we have that $P\left(f_k^O\right) \to \infty$. Our goal in choosing this sequence is to balance the aforementioned two sources of error in order to derive a tight upper bound on the convergence rate of our penalized estimator.

Note that the sequence of representatives $f_0^O, f_1^O, \ldots$ in (ii) is just a tool to prove the rates of convergence. This sequence is not actually used to construct our penalized regression estimator.

2.1. Illustrative example

First, we consider an illustrative example; the theory for this example is discussed in Section 3. Here, we estimate the function $f^* = I(x > 0)$ using functions in $\mathcal{F}_2 = \left\{f \,\middle|\, f \text{ is 2 times weakly differentiable, } \int_x \left|f''(x)\right| dx < \infty\right\}$. Note

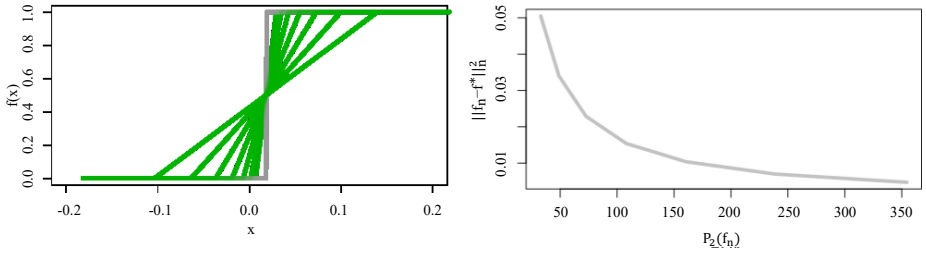

Figure 1. These figures show a sequence of representatives in $\mathcal{F}_2$ used to approximate $f^* = I(x > 0)$. The left figure shows our sequence of *soft indicators*, approximating our original indicator function f^*. I^δ with larger δ are indicated by darker shades of green. On the right, we show the trade-off between $\left\|f^* - I^\delta\right\|_n^2$ on the y-axis, and $P_2\left(I^\delta\right)$ on the x-axis for varying δ.

that $f^* \notin \mathcal{F}_2$ (it is not twice, weakly differentiable), though we do have $f^* \in \mathcal{F}_1$. However, we can approximate f^* using what we will call the *soft indicator function*, $I^\delta \in \mathcal{F}_2$, defined as

$$I^\delta(x) = \begin{cases} 0, & x \leq -\delta, \\ \frac{(x+\delta)_+}{2\delta}, & -\delta \leq x \leq \delta, \\ 1, & \delta \leq x. \end{cases}$$

In the left plot of Figure 1, we can see that as $\delta \to 0$, our soft indicator function visually approximates f^* increasingly well; however, its first derivative changes increasingly sharply ($P_2\left(I^\delta\right) \to \infty$). In particular, the right plot of Figure 1 explicitly shows the trade-off between $\left\|f^* - I^\delta\right\|_n^2$ and $P_2\left(I^\delta\right)$ as δ varies. Given a sequence $\delta_n \to 0$, we can define a sequence of approximators $f_n^O \equiv I^{\delta_n}$.

2.2. Representative inequalities

We now discuss the tools needed to bound the convergence rate for the estimation error of the penalized regression in (1.2) with a penalty P, when f^* lives in the closure of $\mathcal{F}$, but not necessarily in its interior. As noted below, similar results appear elsewhere in the literature, often as intermediate steps in establishing other properties. We present the results here in the minimal form required for our use.

We begin with a *basic inequality* that does not require f^* to be in $\mathcal{F}$. Here, we use a representative $f^O \in \mathcal{F}$. This is similar to the usual basic inequalities used in van de Geer (2000), Bühlmann and van de Geer (2011), and elsewhere. However, with a little extra work, we have $\|\hat{f} - f^O\|_n^2$ on the left-hand-side, which

we need in the proof of the next theorem.

Lemma 1. *Define f^* as in* (1.1) *(potentially not in $\mathcal{F}$), and $\hat{f}$ as in* (1.2) *(in $\mathcal{F}$). Let f^O be any other function in $\mathcal{F}$. Suppose P is convex. Let $\langle h, g\rangle_n = (1/n)\sum_{i=1}^n h(x_i)\, g(x_i)$ for any functions h, g. Then,*

$$\left\|\hat{f} - f^*\right\|_n^2 + \left\|\hat{f} - f^O\right\|_n^2 + 2\lambda P\left(\hat{f}\right) \leq \left\|f^O - f^*\right\|_n^2 + 2\left\langle \epsilon, \hat{f} - f^O\right\rangle_n + 2\lambda P\left(f^O\right). \tag{2.1}$$

The proof of this lemma is given in the online Supplement Material Section S.1, and is a straightforward calculation. Similar ideas appear in the *two-point inequality* and *two point margin* of van de Geer (2016) which are used to establish oracle inequalities in sparse parametric regression, and in Sadhanala and Tibshirani (2017), who focus on additive modeling.

Now, given a sequence of representatives $f_0^O, f_1^O, \ldots \in \mathcal{F}$, with $\left\|f_n^O - f^*\right\|_n^2 \to 0$, we combine an argument in van de Geer (2000) with the basic inequality above to establish an upper bound for the rate of convergence of $\hat{f}$ to f^*.

Before stating our first result, we review the definition of *metric entropy*. For F, a subset of our space $\mathcal{F}$, the covering number with respect to the empirical norm, $N(\delta, F, \|\cdot\|_n)$, is defined as the minimum number of balls of radius δ (with respect to the empirical norm) that are required such that F is a subset of the union of those balls. The *metric entropy* is defined as the log of the covering number: $H(\delta, F, \|\cdot\|_n) \equiv \log N(\delta, F, \|\cdot\|_n)$. Modifying the arguments and results of Theorem 10.2 of van de Geer (2000), we arrive at the following results.

Theorem 1. *Suppose data are generated as in* (1.1)*, and $\hat{f}$ is defined as in* (1.2)*, for some $\lambda_n > 0$. Suppose P is a seminorm, and $\mathcal{F}$ is a linear subspace of $\mathcal{L}_2[-1, 1]$. Let $f_0^O, f_1^O, \ldots \in \mathcal{F}$, with $P\left(f_k^O\right) > 0$ for all k. Suppose the metric entropy of $\mathcal{F}$ has a polynomial bound, given by*

$$H\left(\delta, \{f \in \mathcal{F} \mid P(f) \leq 1\}, \|\cdot\|_n\right) \leq A\delta^{-\alpha}, \tag{2.2}$$

for some $A > 0$ and $\alpha \in (0, 2)$.

If we choose

$$\lambda_n^{-1} = O_p\left(n^{2/(2+\alpha)} P^{(2-\alpha)/(2+\alpha)}\left(f_n^O\right)\right),$$

then,

$$\left\|\hat{f} - f^*\right\|_n^2 \leq \left\|f^* - f_n^O\right\|_n^2 + O_p\left(\lambda_n P\left(f_n^O\right)\right). \tag{2.3}$$

The theorem is proved in the online Supplementary Material, Section S.2. A similar result is given in the recent paper by Sadhanala and Tibshirani (2017) for the more general case of additive models. From here, we can optimize (2.3) over λ_n, choosing $\lambda_n = n^{-2/(2+\alpha)} P^{(-2+\alpha)/(2+\alpha)}\left(f_n^O\right)$. Now, we need to choose our sequence $\left\{f_n^O\right\}$ to balance the two terms on the right-hand-side of (2.3): a term that depends only on the approximation error of our representative sequence,

$$\left\|f^* - f_n^O\right\|_n^2, \tag{2.4}$$

and a term that depends only on the entropy of our class and on the *complexity* (as measured by $P(\cdot)$) of our approximating sequence,

$$\lambda_n P\left(f_n^O\right) = n^{-2/(2+\alpha)} P^{2\alpha/(2+\alpha)}\left(f_n^O\right). \tag{2.5}$$

Note that if f^* is in $\mathcal{F}$ and Eq (2.2) holds for $\mathcal{F}$ with $\alpha < 2$, then (2.5) is the minimax rate for estimating f^* over functions $f \in \mathcal{F}$ with $P(f) \leq P(f^*)$; this is the rate achieved by a penalized regression estimator with a suitable choice of λ_n.

To recap, we have shown that, for $f^* \notin \mathcal{F}$, given any sequence of representatives $f_0, f_1, \ldots \in \mathcal{F}$, we can characterize $\|\hat{f} - f^*\|_n^2$ as the sum of two terms, a misspecification error (2.4), and an estimation error (2.5). The optimal choice of our representative sequence $\left\{f_n^O\right\}$ depends on f^* and the penalty $P(\cdot)$. In the next two sections, we focus on estimating regression functions with bounded higher-order total-variation, and those with Sobolev smoothness.

3. Rates for Bounded Total-Variation Classes

Recall that the kth-order total variation of a function f is defined as

$$P_k(f) \equiv \sup_{x_1,\ldots,x_M} \sum_{m=1}^{M} \left|f^{(k-1)}(x_{m+1}) - f^{(k-1)}(x_m)\right|,$$

where the supremum is taken over all partitions $x_1 < \cdots < x_M$. Let $\mathcal{F}_k$ denote the set of functions with bounded kth-order total variation. Then,

$$\mathcal{F}_k \equiv \{f : [-1,1] \to \mathbb{R} \,|\, f \text{ is } k\text{-times weakly differentiable and } P_k(f) < \infty\}.$$

In this section, we investigate the convergence rate of the penalized estimator in (1.2), with penalty P_k, when the true function f^* is not in $\mathcal{F}_k$, but is in $\mathcal{F}_{l+1}$, for some $l + 1 < k$. Specifically, by specializing our results in Theorem 1, we

establish an upper bound for the convergence rate of the penalized estimator in two cases: $l = 0$ and $l \geq 1$.

As a reminder, if we had assumed that f^* was in $\mathcal{F}_{l+1}$, for some $l + 1 \geq k$ (rather than $<$), then, in fact, we would have $f^* \in \mathcal{F}_k$, and the estimator would converge at the minimax rate over any bounded subset of $\mathcal{F}_k$.

For the results in this section, we assume there exists a universal L, such that $\|f^*\|_\infty \leq L$. We further assume that problem (1.2) is solved under the constraint that $\|\hat{f}\|_\infty \leq L$. This is not necessary to prove our results, but greatly eases the exposition. In addition we assume $\mathcal{X} = [-1, 1]$, and that we have a fixed design, with evenly spaced x_i. In particular, define the triangular array $x_{i,n} = 2i/n - 1$, for $i \leq n$, and slightly alter the definition of our empirical norm: $\|g\|_n^2 \equiv (1/n) \sum g(x_{i,n})^2$. The following results can be shown for a random design. In particular, in the case of Sobolev and total-variation-type penalties, the entropy bounds in (2.2) hold with respect to $\|\cdot\|_\infty$, rather than just $\|\cdot\|_n$ (Nickl and Pötscher (2007)). Thus, for any x, the entropy bounds in (2.2) hold, making the transition to stochastic x relatively straightforward.

3.1. Choice of λ

Recall that, in practice, nonparametric functions are often estimated from observed data by solving the penalized regression problem in (1.2) for a given choice of P, with λ chosen using, for example, cross-validation. The results in Lemma 2 and Lemma 3, give us the rates of convergence if we choose an oracle λ-value (rather than selecting λ using cross-validation). In a number of cases, these λ-values depend on both k and l, which are unknown, and thus cannot be substituted into (1.2) for estimation. However, using recent ideas (Feng and Simon (In Press)), these results can still be useful in understanding the performance of a penalized regression estimator with λ selected by split-sample validation. We discuss this further in Section S.7 of the online Supplementary Material. We hope to engage with this further in future work.

Note that the "approximating sequences" we discuss below are technical tools we use to show these rates; they are not used in the estimation procedure.

3.2. Estimating $f^* \in \mathcal{F}_1$ using P_k

We first consider estimating $f^* \in \mathcal{F}_1$ using the penalty P_k, for $k > 1$. We obtain the following result,

Lemma 2. *Suppose $f^* \in \mathcal{F}_1$, $\hat{f}$ is given by solving* (1.2)*, using $P(\cdot) = P_k(\cdot)$ for*

$k > 1$, and $\lambda_n = n^{-2k^2/(4k-1)} P_1 (f^*)^{-(2k-1)^2/(4k-1)}$. Then,

$$\left\| \hat{f} - f^* \right\|_n^2 = O_p \left(n^{-2k/(4k-1)} P_1 (f^*)^{(4k-2)/(4k-1)} \right). \tag{3.1}$$

The main idea here is to use a two-stage approximation. First, we note that any $f^* \in \mathcal{F}_1$ can be approximated by a piecewise constant function $\tilde{f}_n(x) = \sum_{j=1}^{J(n)} \beta_{j,n} I(x > d_{j,n})$ with $J(n)$ knots, $d_{1,n}, \ldots, d_{J(n),n}$ (Birman and Solomyak (1967)). However, $\tilde{f}_n \notin \mathcal{F}_k$. Thus, we also approximate the indicator function $I(x > 0)$ using what we refer to as the *kth-order soft indicator function*:

$$I_k^\delta(x) \equiv \delta^{-1} \int_{-\infty}^{x} b_{k-1} \left(\frac{t}{\delta} \right) dt.$$

Here, b_{k-1} denotes the cardinal b-spline of order $k-1$, scaled to have support on $[-1, 1]$. Note that b_{k-1} is a piecewise $k-1$-order polynomial, which is nonnegative and integrates to one(Udovičić (2010)). Importantly, this soft indicator is an element of $\mathcal{F}_k$. Therefore, our final representative becomes

$$f_n(x) = \sum_{j=1}^{J(n)} \beta_{j,n} I_k^{\delta_n} (x - d_{j,n}).$$

From here, we can bound $\|f_n - f^*\|_n^2$ and $P_k(f_n)$ as functions of δ_n, and select a suitable δ_n to obtain the rates in (3.1). The details of the proof are given in the online Supplementary Material Section S.3.1.

3.3. Estimating $f^* \in \mathcal{F}_{l+1}$ using P_k for $1 < l + 1 < k$

We now give a similar result for estimating a regression function with $(l+1)$-st order bounded variation, using a penalized regression with penalty P_k for $k > l + 1 > 1$. In this case, we get:

Lemma 3. *Let $k > l + 1 \geq 2$. Suppose $f^* \in \mathcal{F}_{l+1}$, $\hat{f}$ is given by solving (1.2), using $P(\cdot) = P_k(\cdot)$, and $\lambda_n = n^{-k(k+1-l)/(3k-l)} P_{l+1}(f^*)^{(1-2k)(k-l)/(3k-l)}$. Then, we have*

$$\left\| \hat{f} - f^* \right\|_n^2 = O_p \left(n^{-2k/(3k-l)} P_{l+1} (f^*)^{(2k-2l)/(3k-l)} \right). \tag{3.2}$$

Lemma 3 is not a generalization of Lemma 2. Here, we require $l + 1 \geq 2$, which allows us to use differentiability to obtain a slightly faster rate.

The crux of the argument is similar to the $\mathcal{F}_1$ case. We again use a two-stage approximation, but our approximating sequence is different. Using a similar

result to that in Section 3.2, $f^* \in \mathcal{F}_{l+1}$ can be approximated by an lth-order spline: $\tilde{f}_n(x) = \sum_{j=1}^{J(n)} \beta_{j,n} (x - d_{j,n})_+^l$ (Birman and Solomyak (1967)). However, $\tilde{f}_n$ is not in $\mathcal{F}_k$. To derive an estimate that is in $\mathcal{F}_k$, we approximate x_+^l using integrals of our kth-order soft-indicator function. In particular, we consider:

$$\psi_{k,l}^{\delta}(x) \equiv l!\, \delta^{-1} \underbrace{\int_{-\infty}^{x} \cdots \int_{-\infty}^{t_2}}_{(l+1) \text{ times}} b_{k-l-1}\left(\frac{t_1}{\delta}\right) dt_1 \cdots dt_{l+1},$$

which is an element of $\mathcal{F}_k$ that closely approximates x_+^l. The approximation becomes better as $\delta \to 0$. In addition, note that $\partial^l \psi_{k,l}^{\delta} / \partial x^l = l! I_{k-l}^{\delta}$, which, in particular, closely approximates $\partial^l \left(x_+^l\right) / \partial x^l = l! I(x \geq 0)$. Our final representative becomes

$$f_n(x) = \sum_{j=1}^{J(n)} \beta_{j,n} \psi_{k,l}^{\delta_n} (x - d_{j,n}).$$

By selecting suitable δ_n we obtain our rates in (3.2). The details of the proof are given in online Supplementary Material Section S.3.5.

3.4. A comparison of rates

The minimax rate for estimating a function in $\mathcal{F}_{l+1}$ is $n^{-(2l+2)/(2l+3)}$. For $l = 0$, this gives $n^{-2/3}$. In comparison, our results show that using P_k with $k > 1$ yields a rate of at least $n^{-2k/(4k-1)}$. This is substantially slower than the minimax rate. However, it is always faster than $n^{-1/2}$, which is the best rate one can achieve using a linear smoother when estimating a function from $\mathcal{F}_1$ (Donoho et al. (1995)). The $n^{-1/2}$ rate is also the critical rate needed if the nonparametric estimate of the regression function is used as an intermediate quantity when estimating a path-wise differentiable functional (Bickel et al. (1998)). As long as the regression function is estimated at a rate faster than $n^{-1/2}$, we can be efficient in our downstream estimate of that functional, and apply semi-parametric tools to build asymptotically valid confidence intervals.

For estimating a regression function in $\mathcal{F}_{l+1}$, for $l \geq 1$, our rate of $n^{-2k/(3k-l)}$ is not too far from the optimal $n^{-(2l+2)/(2l+3)}$. It is always faster than $n^{-2/3}$, which is the rate one would achieve using P_1 for functions in $\mathcal{F}_{l+1}$. In addition, as $k, l \to \infty$, with $(k-l)/k \to 0$, our rate converges to the parametric rate n^{-1}.

The empirical results of Section 5 lead us to conclude that our results in Lemma 2, for the $\mathcal{F}_1$ case, are sharp (or nearly sharp). However, the results in Lemma 3, for $l + 1 \geq 2$, do not appear to be sharp. We believe that this is

because our approximation $\psi^{\delta}_{k,l}(x)$ only agrees with $\psi_l \equiv (x)^l_+$ in its lth derivative at $x = \delta$; for instance, we do not even have $\psi^{\delta}_{k,l}(\delta) = \psi_l(\delta)$. Ideally, we would have $\psi^{\delta}_{k,l}(\delta) = \psi_l(\delta)$ and $\partial^m \psi^{\delta}_{k,l}(\delta)/\partial x^m = \partial^m \psi_l(\delta)/\partial x^m$, for all $m \leq l$, or, equivalently, an approximator with $\psi^{\delta}_{k,l}(x) = \psi_l(x)$, for all $x \geq \delta$. If such an approximator could be found, then the misspecification error would become δ^{2l+1} (instead of δ^2). If, in addition, $P_k(\psi^{\delta}_{k,l})$ scales as $1/\delta^{k-l-1}$, then our convergence rate in Lemma 3 would improve to $n^{-2k(l+1)/(2k(2l+2)-1)}$. This rate matches the empirical results of Section 5.

4. Rates for Sobolev Classes

Many popular nonparametric procedures assume that f^* has finite a Sobolev semi-norm

$$P^d_k(f) = \left(\int \left| f^{(k)}(x) \right|^d dx \right)^{1/d},$$

where, classically, d is taken to be two (Hastie and Tibshirani (1990); Craven and Wahba (1978)). We define

$$\mathcal{F}^d_k = \left\{ f : [-1,1] \to \mathbb{R} \,\middle|\, f \text{ is } k\text{-times differentiable and } P^d_k(f) < \infty \right\}.$$

In this section, we explore the convergence rates of the estimators obtained a) using Sobolev semi-norms (P^d_k, $d > 1$) as penalties, and b) when f^* lies in a Sobolev space $\mathcal{F}^d_l$, for $d > 1$, rather than in a space of bounded total-variation. As before, we are interested in cases where, for our given choice of penalty (Sobolev or total-variation semi-norm), we have $P(f^*) = \infty$.

First, note that $\mathcal{F}^d_k \subset \mathcal{F}_k$, for all $d \geq 1$. This follows from the simple L_d-norm inequality

$$\left(\int \left| f^{(k)}(x) \right|^d dx \right)^{1/d} \geq \int \left| f^{(k)}(x) \right| dx. \tag{4.1}$$

In particular, this immediately gives us the following results when using a kth-order total-variation penalty to estimate a function that actually lies in a lower order Sobolev class,

Corollary 1. *For $d > 1$, suppose $f^* \in \mathcal{F}^d_1$, $\hat{f}$ is found by solving (1.2), using $P(\cdot) = P_k(\cdot)$ for $k > 1$, and $\lambda_n = n^{-2k^2/(4k-1)} P_1(f^*)^{-(2k-1)^2/(4k-1)}$. Then,*

$$\left\| \hat{f} - f^* \right\|^2_n = O_p\left(n^{-2k/(4k-1)} P_1(f^*)^{(4k-2)/(4k-1)} \right). \tag{4.2}$$

Corollary 2. *Let $k > l+1 \geq 2$ and $d > 1$. Suppose $f^* \in \mathcal{F}^d_{l+1}$, and $\hat{f}$ is found by solving (1.2), using $P(\cdot) = P_k(\cdot)$, and $\lambda_n = n^{-k(k+1-l)/(3k-l)} P_{l+1}(f^*)^{(1-2k)(k-l)/(3k-l)}$. Then,*

$$\left\|\hat{f} - f^*\right\|_n^2 = O_p\left(n^{-2k/(3k-l)} P_{l+1}\left(f^*\right)^{(2k-2l)/(3k-l)}\right). \tag{4.3}$$

Note that the total-variation norms $P_1(f^*)$ in Corollary 1 and $P_{l+1}(f^*)$ in Corollary 2 are finite, owing to (4.1); these can be replaced with $P^d_1(f^*)$ and $P^d_{l+1}(f^*)$, respectively; however, that would give a looser bound.

It is also of interest to know how a penalized estimator using a higher-order Sobolev penalty performs when the true f^* lies in a lower order Sobolev or total-variation class. We begin by considering the use of P^d_k if f^* is truly only in P_l, for some $l \leq k$. We obtain the following results.

Lemma 4. *Suppose $f^* \in \mathcal{F}_1$, $\hat{f}$ is found by solving (1.2), using $P(\cdot) = P^d_k(\cdot)$ for $k > 1$ and $d > 1$, and $\lambda_n = n^{-(2k^2+4k)/(4k+1)} P_1(f^*)^{-(4k^2-1)/(4k+1)}$. Then,*

$$\left\|\hat{f} - f^*\right\|_n^2 = O_p\left(n^{-2k/(4k+1)} P_1\left(f^*\right)^{(4k+2)/(4k+1)}\right). \tag{4.4}$$

Lemma 5. *Let $k > l+1 \geq 2$. Suppose $f^* \in \mathcal{F}_{l+1}$, $\hat{f}$ is given by solving (1.2), using $P(\cdot) = P^d_k(\cdot)$ for $d > 1$, and $\lambda_n = n^{-k(k+2-l)/(3k-l+1)} P_{l+1}(f^*)^{-(2k(k+2-2l)+1)/(3k-l+1)}$. Then, we have*

$$\left\|\hat{f} - f^*\right\|_n^2 = O_p\left(n^{-2k/(3k-l+1)} P_{l+1}\left(f^*\right)^{(2k-2l)/(3k-l+1)}\right). \tag{4.5}$$

The proofs of these lemmas are given in the online Supplementary Material Sections S.4 and S.5. Note, that we use $\psi^\delta_{k+1,l}(x) \in \mathcal{F}^d_k$ to approximate $x^l I(x \geq 0)$. This differs from the situation in Section 3.3, where we could use $\psi^\delta_{k,l}(x) \in \mathcal{F}_k$; in this case, however, $\psi^\delta_{k,l}(x) \notin \mathcal{F}^d_k$ for $d > 1$. This is largely responsible for the slightly worse rate than those of Lemmas 2, and 3.

By again using the seminorm inequality in (4.1), we can immediately extend the results in Lemmas 4 and 5 to the case where f^* lies in a lower-order Sobolev class.

Corollary 3. *Suppose $f^* \in \mathcal{F}^d_1$, $\hat{f}$ is found by solving (1.2), using $P(\cdot) = P^d_k(\cdot)$ for $k > 1$ and $d > 1$, and $\lambda_n = n^{-(2k^2+4k)/(4k+1)} P_1(f^*)^{-(4k^2-1)/(4k+1)}$. Then,*

$$\left\|\hat{f} - f^*\right\|_n^2 = O_p\left(n^{-2k/(4k+1)} P_1\left(f^*\right)^{(4k+2)/(4k+1)}\right). \tag{4.6}$$

Corollary 4. *Let* $k > l+1 \geq 2$. *Suppose* $f^* \in \mathcal{F}^d_{l+1}$, $\hat{f}$ *is found by solving* (1.2), *using* $P(\cdot) = P^d_k(\cdot)$ *for* $d > 1$, *and* $\lambda_n = n^{-k(k+2-l)/(3k-l+1)} P_{l+1}(f^*)^{-(2k(k+2-2l)+1)/(3k-l+1)}$. *Then,*

$$\left\| \hat{f} - f^* \right\|_n^2 = O_p\left(n^{-2k/(3k-l+1)} P_{l+1}\left(f^*\right)^{(2k-2l)/(3k-l+1)} \right). \tag{4.7}$$

Once again, these results hold with $P_{l+1}(f^*)$ and $P_1(f^*)$ replaced with $P^d_{l+1}(f^*)$ and $P^d_1(f^*)$, respectively. However, the bounds are looser.

As in the case of Lemma 3, we do not believe the results in Lemma 5 or those in Corollaries 2 and 4 are tight. As discussed in Section 3.4, an approximating spline with more matching derivatives outside our shrinking interval, would immediately obtain a faster rate.

5. Simulation

To empirically evaluate the sharpness (or lack thereof) of the rates shown in Section 3, we ran a set of simulation experiments. For varying values of n, we used $x_i = \{1/n, 2/n, \ldots, 1\}$, and generated y_i as

$$y_i = f^*(x_i) + \epsilon_i,$$

where ϵ_i are independent and identically distributed (i.i.d.) and from various distributions (Gaussian, uniform, and double-exponential). We used piecewise-constant and piecewise-linear functions for f^*. In this section we present the results for ϵ_i drawn from $N(0,1)$; and f^* piecewise-constant and piecewise-linear with a single knot. Simulations using other settings are given in Section S.8 of the online Supplementary Material.

Here we use the piecewise constant $f^*_1(x) \equiv 3 * I(x > 0.5)$ to evaluate the rate in Lemma 2, and the piecewise linear $f^*_2(x) = 3(x-0.5)_+$ for Lemma 3. We used $n_j = 200 * 1.5^j$ for $j = 1, \ldots, 6$, and ran 400 simulations for each n. In each simulation, we evaluated the MSE of the penalized regression with P_1, P_2, and an approximation to P_3, known as ℓ_1 trend filtering (Tibshirani (2014)); we also considered the oracle model, where we fit a parametric model with a zero-th-order or first-order spline, with a single knot at 0.5. Unfortunately, an exact fit for P_3 is computationally difficult. However, the quadratic trend filtering estimator that we use can be fit efficiently using the method of Ramdas and Tibshirani (2016). This is also a penalized estimator and has been shown to be asymptotically equivalent to a penalized regression with penalty P_3 (Tibshirani

Table 1. Estimated slope (and standard error) of the regression of log(MSE) on log(n) for our estimator based on total variation penalty, as well as an oracle estimator. In this example, slopes were calculated based on 400 simulated experiments with each of $n_j = 200{*}1.5^j$ ($j = 1, \ldots, 6$). The results in the first column correspond to data generated from $f_1^*(x) \equiv 3{*}I(x > 0.5)$; those in the second column are based on $f_2^*(x) = 3(x-0.5)_+$.

Estimator/Penalty	slope for f_1^*	slope for f_2^*
P_1	-0.831 (0.021)	-0.656 (0.012)
P_2	-0.572 (0.006)	-0.863 (0.025)
P_3	-0.548 (0.005)	-0.780 (0.018)
oracle	-0.967 (0.036)	-0.943 (0.037)

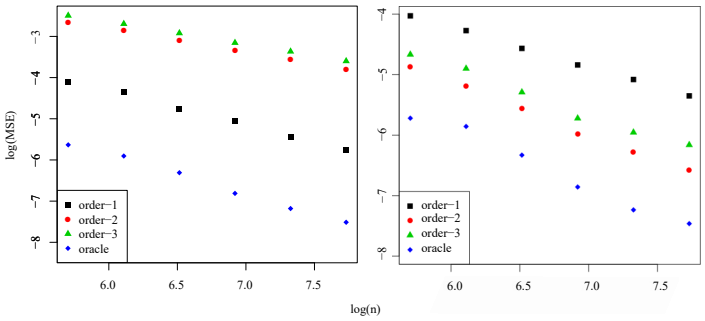

Figure 2. Average log(MSE) vs. log(n) for estimators with total-variation penalties of degree 1, 2, and 3, along with a parametric oracle. In the left panel, data were generated using the regression function $f^*(x) = 3{*}I(x > 0.5)$; in the right panel, $f^*(x) = 3(x - 0.5)_+$ was used. The MSE was calculated as the average over 400 simulations for each $n_j = 200 * 1.5^j$, for $j = 1, \ldots, 6$.

(2014)). These estimates were all fit using the `R` package `glmgen`. To select λ, we used an oracle procedure. For each penalty/simulation, we tried a range of λ-values, and selected the value that minimized the MSE, $\|f^* - \hat{f}_\lambda\|_n^2$. We then reported the MSE of the estimator with that optimal λ-value.

Table 1 and Figure 2 show the results of estimating the piecewise constant and linear regression functions, f_1^* and f_2^*, respectively. Here, we regress log(MSE) on log(n). The slope of this regression provides an estimate of the exponent in our convergence rates. In the first column of Figure 1, based on our theory, we would expect to see a slope of $-4/7 = -0.571$ when using P_2 (rate for $\mathcal{F}_1$ with $k = 2$), and $-6/11 = -0.545$ when using P_3 (rate for $\mathcal{F}_1$ with $k = 3$). The rates from our simulations are very close to these theoretical rates. These findings suggest that our bound in Lemma 2 is sharp. Note that P_1 achieves a faster rate than the minimax $-2/3$ over $\mathcal{F}_1$. This is because

the penalized estimator using a kth-order total-variation penalty can adapt and achieve a near-parametric rate when estimating a function that is a piecewise $(k-1)$-degree polynomial (Guntuboyina et al. (2017)). In the second column, we expect to see $-2/3$ when using P_1 (the minimax rate for functions of bounded first-order TV), and a rate of at least $-3/4$ using P_3 (rate for $l+1=2$, with $k=3$). The simulations appear to suggest a slightly faster rate for P_3, indicating that the upper bound in Lemma 3 may not be sharp. In fact, our empirical rate matches our hypothesized convergence rate from Section 3.4: For $k=3$ and $l=1$, the hypothesized rate reduces to $-18/23 \approx -0.782$. This is the rate our bound would obtain if we could find a slightly better behaved approximating sequence. Finally, using P_2, we see an empirical rate that is slightly faster than the minimax rate over $\mathcal{F}_2$; this, again, is in line with the work of Guntuboyina et al. (2017).

6. Discussion

W have discussed a framework for proving the convergence rates of penalized regression estimators under misspecified smoothness, when $P(f^*) = \infty$. In this framework, the error from a penalized regression comprises two parts: (a) the distance from f^O, any representative in $\mathcal{F}$, to f^*; and (b) the rate at which we could have estimated f^O if it were the true conditional mean; the latter rate involves a term $P\left(f^O\right)$. We applied this framework to the estimation of functions with bounded $(l+1)$st-order total-variation or Sobolev variation, using a $k > l+1$ order total-variation, or Sobolev penalty. We provide a bound on the convergence rate, and show, in particular, that when using a total-variation penalty, for any l, k the rate is faster than $n^{-1/2}$. The $n^{-1/2}$ rate is the critical rate needed if the nonparametric estimate of the regression function is to be used as an intermediate quantity for estimating a path-wise differentiable functional (Bickel et al. (1998)). As long as the regression function is estimated at a rate faster than $n^{-1/2}$, we can in general be efficient in our downstream estimate of that functional, and apply semi-parametric tools to build asymptotically valid confidence intervals. Applications of these flexible methods are becoming increasingly popular in causal inference (Chernozhukov et al. (2016)), where it is critical to guarantee that the estimators of the nuisance parameters obtain this rate.

We conducted a simulation experiment to evaluate the sharpness of our bounds using total-variation penalties. The results show that our bound in

Lemma 2 appears sharp, whereas that in Lemma 3 does not. We suspect that this is because of the approximating sequence we chose. We can obtain a tighter bound by constructing an approximating sequence that matches on lower-order derivatives.

The proposed framework can be applied more generally. For example, using the results from Sadhanala and Tibshirani (2017), it would be straightforward to derive similar rates for additive models. In addition, it might be of interest to consider general multivariate penalized smoothers.

Supplementary Material

The online Supplementary Material contains formal proofs of all results (lemmas and theorems) in the manuscript, an in-depth discussion of the tuning parameter λ, and additional simulation results.

References

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory* **39**, 930–945.

Bickel, P. J., Klaassen, C. A., Ritov, Y. and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag.

Birman, M. S. and Solomyak, M. Z. (1967). Piecewise-polynomial approximations of functions of the classes W_p^α. *Matematicheskii Sbornik* **115**, 331–355.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, Berlin.

Candes, E. J. (1998). *Ridgelets: Theory and Applications*. Ph.D. thesis, Stanford University Stanford.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. and Newey, W. K. (2016). Double machine learning for treatment and causal parameters. Technical report, cemmap working paper, Centre for Microdata Methods and Practice.

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik* **31**, 377–403.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90**, 1200–1224.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **57**, 301–369.

Feng, J. and Simon, N. (In Press). An analysis of the cost of hyper-parameter selection via split sample validation, with applications to penalized regression. *Statistica Sinica* **30**, 511–530.

Guntuboyina, A., Lieu, D., Chatterjee, S. and Sen, B. (2017). Spatial adaptation in trend filtering. *arXiv preprint arXiv:1702.05113* .

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Wiley Online Library.

Lepski, O. V. and Spokoiny, V. G. (1997). Optimal pointwise adaptive methods in nonparametric

estimation. *The Annals of Statistics* **25**, 2512–2546.

Mammen, E. and van de Geer, S. (1997). Locally adaptive regression splines. *The Annals of Statistics* **25**, 387–413.

Nickl, R. and Pötscher, B. M. (2007). Bracketing metric entropy rates and empirical central limit theorems for function classes of besov-and sobolev-type. *Journal of Theoretical Probability* **20**, 177–199.

Omranian, N., Eloundou-Mbebi, J. M., Mueller-Roeber, B. and Nikoloski, Z. (2016). Gene regulatory network inference using fused Lasso on multiple data sets. *Scientific Reports* **6**, 20533.

Ramdas, A. and Tibshirani, R. J. (2016). Fast and flexible admm algorithms for trend filtering. *Journal of Computational and Graphical Statistics* **25**, 839–858.

Saâda-Bouzid, E., Defaucheux, C., Karabajakian, A., Coloma, V. P., Servois, V., Paoletti, X., Even, C., Fayette, J., Guigay, J., Loirat, D., et al. (2017). Hyperprogression during anti-pd-1/pd-l1 therapy in patients with recurrent and/or metastatic head and neck squamous cell carcinoma. *Annals of Oncology* **28**, 1605–1611.

Sadhanala, V. and Tibshirani, R. J. (2017). Additive models with trend filtering. *arXiv preprint arXiv:1702.05037* .

Starck, J.-L., Candès, E. J. and Donoho, D. L. (2002). The curvelet transform for image denoising. *IEEE Transactions on Image Processing* **11**, 670–684.

Szczesniak, R. D., McPhail, G. L., Duan, L. L., Macaluso, M., Amin, R. S. and Clancy, J. P. (2013). A semiparametric approach to estimate rapid lung function decline in cystic fibrosis. *Annals of Epidemiology* **23**, 771–777.

Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics* **42**, 285–323.

Udovičić, Z. (2010). Splines in numerical integration. *Mathematica Balkanica New Series* **24**, 351–358.

van de Geer, S. (2000). *Empirical Processes in M-estimation*, volume 6. Cambridge University Press, New York.

van de Geer, S. (2016). *Estimation and Testing Under Sparsity*. Springer, Switzerland.

Noah Simon

1705 Northeast Pacific Street University of Washington, Box 357232 Seattle, WA 98195, USA.

E-mail: nrsimon@uw.edu

Ali Shojaie

1705 Northeast Pacific Street University of Washington, Box 357232 Seattle, WA 98195, USA.

E-mail: ashojaie@uw.edu

(Received April 2018; accepted April 2019)