Convergence Rates of Nonparametric Penalized Regression under Misspecified Smoothness

Noah Simon and Ali Shojaie

University of Washington, Department of Biostatistics

Supplementary Material

S1. Proof of Lemma 2.1

To prove this lemma we will first state and prove two short sub-lemmas:

Lemma S1.1. *Suppose $\hat{f}$, f^* and f^O are functions that map to $\mathbb{R}$. Then,*

$$2\left\langle \hat{f} - f^*, \hat{f} - f^O \right\rangle_n = \left\|\hat{f} - f^*\right\|_n^2 + \left\|\hat{f} - f^O\right\|_n^2 - \left\|f^* - f^O\right\|_n^2$$

The proof follows from arithmetic. This lemma is not new, and has been recently used to analyze the MSE of misspecified models in a parametric context [26]. This can be thought of as a generalized law of cosines.

Lemma S1.2. *If $\hat{f}$ is defined according to (12), and $f^O \in \mathcal{F}$ is any other* *function, then*

$$\left\langle y - \hat{f}, f^O - \hat{f} \right\rangle_n \leq 2\lambda P\left(f^O\right) - 2\lambda P\left(\hat{f}\right)$$

Proof. The proof follows from the KKT conditions. For $\epsilon \in [0, 1]$ define $f_\epsilon = \hat{f} + \epsilon\left(f^O - \hat{f}\right)$. Because $\mathcal{F}$ is convex, $f_\epsilon \in \mathcal{F}$ for all $\epsilon \in [0, 1]$. Now, let's consider the one dimensional problem

$$\hat{\epsilon} \equiv \operatorname{argmin}_{\epsilon\in[0,1]} \frac{1}{2} \|y - f_\epsilon\|_n^2 + \lambda P\left(f_\epsilon\right). \tag{S1.1}$$

Because $\hat{f}$ minimizes (12), we know $\hat{\epsilon} = 0$ minimizes (S1.1). Thus, since the objective is convex, 0 must be in the sub-differential of the objective in (S1.1) evaluated at $\epsilon = 0$. Taking the sub-gradient at $\epsilon = 0$ we get

$$0 = -\left\langle y - \hat{f}, f^O - \hat{f} \right\rangle_n + \lambda \left\langle \dot{P}\left(\hat{f}\right), f^O - \hat{f} \right\rangle, \tag{S1.2}$$

for some sub-gradient $\dot{P}\left(\hat{f}\right)$ of $P(f_\epsilon)$ evaluated at $\epsilon = 0$. Now by the definition of a sub-gradient we know that $\left\langle \dot{P}\left(\hat{f}\right), f^O - \hat{f} \right\rangle \leq P\left(f^O\right) - P\left(\hat{f}\right)$. Plugging this into (S1.2) we get

$$\left\langle y - \hat{f}, f^O - \hat{f} \right\rangle_n \leq \lambda P\left(f^O\right) - \lambda P\left(\hat{f}\right).$$

□

Now we combine these results to prove Lemma 2.1.

Proof of Lemma 2.1. We begin by using the result of Lemma S1.1

$$\left\|\hat{f} - f^*\right\|_n^2 + \left\|\hat{f} - f^O\right\|_n^2 = 2\left\langle \hat{f} - f^*, \hat{f} - f^O \right\rangle_n + \left\|f^* - f^O\right\|_n^2.$$

Now, we can continue with the first term on the RHS by remembering that $y = f^*(x) + \epsilon$, and then applying Lemma S1.2

$$\begin{aligned} 2\left\langle \hat{f} - f^*, \hat{f} - f^O \right\rangle_n &= 2\left\langle y - \hat{f}, f^O - \hat{f} \right\rangle_n + 2\left\langle \epsilon, \hat{f} - f^O \right\rangle_n \\ &\leq 2\left\langle \epsilon, \hat{f} - f^O \right\rangle_n + 2\lambda P\left(f^O\right) - 2\lambda P\left(\hat{f}\right). \end{aligned}$$

Putting things together, we get

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

as desired. □

S2. Proof of theorem 2.2

We first note that since

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

then this same bound holds (up to a different constant) for normalized functions $\frac{f - f_n^O}{P(f) + P(f_n^O)}$, for any $f_n^O \in \mathcal{F}$, whenever $P(f) + P(f_n^O) > 0$. This is because $f - f_n^O \in \mathcal{F}$, and P a semi-norm, along with $P(f_n^O) > 0$ implies $\frac{P(f - f_n^O)}{P(f) + P(f_n^O)} \leq 1$. Thus,

$$H\left(\delta, \left\{\frac{f - f_n^O}{P(f) + P(f_n^O)} \,\middle|\, f \in \mathcal{F}\right\}, \|\cdot\|_n\right) \leq A\delta^{-\alpha},$$

for all $\delta > 0, n \geq 1$. Now, given any $\epsilon > 0$, using Lemma 8.4 of van de Geer [25] we have that

$$\sup_{f\in\mathcal{F}} \frac{\left|\left\langle \epsilon, f - f_n^O\right\rangle_n\right|}{\|f - f_n^O\|_n^{1-\alpha/2}\left(P(f) + P(f_n^O)\right)^{\alpha/2}} \leq C_\epsilon n^{-\frac{1}{2}}. \tag{S2.1}$$

with probability at least $1 - \epsilon$ (where C_ϵ depends only on ϵ). But from Lemma 2.1, we have

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

Plugging (S1.3) in here, we get

$$\begin{aligned}\left\|\hat{f} - f^*\right\|_n^2 + \left\|\hat{f} - f_n^O\right\|_n^2 + 2\lambda_n P\left(\hat{f}\right) \leq{}& \left\|f_n^O - f^*\right\|_n^2 \\ &+ C_\epsilon n^{-\frac{1}{2}}\left\|\hat{f} - f_n^O\right\|_n^{1-\alpha/2}\left(P(\hat{f}) + P(f_n^O)\right)^{\alpha/2} \\ &+ 2\lambda_n P\left(f_n^O\right).\end{aligned} \tag{S2.2}$$

Now, from Young's inequality ($ab \leq a^p/p + b^q/q$ for $1/p + 1/q = 1$) with $p = 4/(2-\alpha)$, and $q = 4/(2+\alpha)$, we get

$$C_\epsilon n^{-\frac{1}{2}}\left\|\hat{f} - f_n^O\right\|_n^{1-\alpha/2}\left(P(\hat{f}) + P(f_n^O)\right)^{\alpha/2} \leq \left\|\hat{f} - f_n^O\right\|_n^2 + \tilde{C}_\epsilon n^{-\frac{2}{2+\alpha}}\left(P(\hat{f}) + P(f_n^O)\right)^{\frac{2\alpha}{2+\alpha}}$$

for some $\tilde{C}_\epsilon$. Plugging this in to (S1.4) we get

$$\left\|\hat{f} - f^*\right\|_n^2 + 2\lambda_n P\left(\hat{f}\right) \leq \left\|f_n^O - f^*\right\|_n^2 + \tilde{C}_\epsilon n^{-\frac{2}{2+\alpha}}\left(P(\hat{f}) + P(f_n^O)\right)^{\frac{2\alpha}{2+\alpha}} + 2\lambda_n P\left(f_n^O\right). \tag{S2.3}$$

We will break the remainder of the argument into two cases: $P\left(\hat{f}\right) \leq P(f_n^O)$ and $P\left(\hat{f}\right) > P(f_n^O)$.

If $P\left(\hat{f}\right) \leq P(f_n^O)$, then (S1.5) reduces to

$$\begin{aligned}\left\|\hat{f} - f^*\right\|_n^2 &\leq \left\|f_n^O - f^*\right\|_n^2 + \tilde{C}_\epsilon n^{-\frac{2}{2+\alpha}} \left(2P(f_n^O)\right)^{\frac{2\alpha}{2+\alpha}} + 2\lambda_n P\left(f_n^O\right) \\ &= \left\|f_n^O - f^*\right\|_n^2 + O_p\left(\lambda_n P\left(f_n^O\right)\right),\end{aligned}$$

where the last line follows because $n^{-\frac{2}{2+\alpha}} P\left(f_n^O\right)^{\frac{2\alpha}{2+\alpha}} = O_p\left[\lambda_n P\left(f_n^O\right)\right]$ by the definition of λ_n.

If instead, $P\left(\hat{f}\right) > P(f_n^O)$, then, we can apply Young's inequality with $p = \frac{2+\alpha}{2\alpha}$ and $q = \frac{2+\alpha}{2-\alpha}$ to the right-hand-side of (S1.5) to get

$$\begin{aligned}\tilde{C}_\epsilon n^{-\frac{2}{2+\alpha}} \left(P(\hat{f}) + P(f_n^O)\right)^{\frac{2\alpha}{2+\alpha}} &\leq \tilde{C}'_\epsilon \left(n^{-\frac{2}{2+\alpha}} \left(2\lambda_n\right)^{\frac{-2\alpha}{2+\alpha}}\right)^{\frac{2+\alpha}{2-\alpha}} + 2\lambda_n\left(P(\hat{f}) + P(f_n^O)\right) \\ &\leq \tilde{C}''_\epsilon \left(n^{\frac{2}{\alpha-2}} \lambda_n^{\frac{2\alpha}{\alpha-2}}\right) + 2\lambda_n\left(P(\hat{f}) + P(f_n^O)\right) \\ &\leq \tilde{C}''_\epsilon n^{-\frac{2}{2+\alpha}} P(f_n^O)^{\frac{2\alpha}{2+\alpha}} + 2\lambda_n\left(P(\hat{f}) + P(f_n^O)\right).\end{aligned}$$

Plugging this into (S1.5), we get

$$\left\|\hat{f} - f^*\right\|_n^2 \leq \left\|f_n^O - f^*\right\|_n^2 + \tilde{C}''_\epsilon n^{-\frac{2}{2+\alpha}} P\left(f_n^O\right)^{\frac{2\alpha}{2+\alpha}} + 4\lambda_n P\left(f_n^O\right).$$

Again, noting that $n^{-\frac{2}{2+\alpha}} P\left(f_n^O\right)^{\frac{2\alpha}{2+\alpha}} = O_p\left[\lambda_n P\left(f_n^O\right)\right]$ (by definition of λ_n), gives us

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

This completes the proof.

S3. Details for Estimating Classes with bounded l-th order TV

Our eventual goal in this section is to characterize the convergence rate obtained using the penalized estimator (12) with penalty P_k when the true function f^* is not in $\mathcal{F}_k$, but is in $\mathcal{F}_{l+1}$ for some $l+1 < k$. In building up to this, and illustrating our method of proof, we give bounds on rates of convergence in the following illustrative examples:

1. Estimating a function in $\mathcal{F}_1$ using P_k $(k > 1)$:
   (a) Piecewise constant function with one knot.
   (b) Piecewise constant function with multiple knots.
   (c) Arbitrary function in $\mathcal{F}_1$.
2. Estimating a function in $\mathcal{F}_{l+1}$ using P_k $(k > l+1 \geq 2)$:
   (a) l-th order spline with one knot.
   (b) l-th order spline with multiple knots.
   (c) Arbitrary function in $\mathcal{F}_{l+1}$.

S3.1 Estimating a function in $\mathcal{F}_1$, using P_k

In this section, we prove Lemma 3.1, giving an upper bound on the rate for estimating a function $f^* \in \mathcal{F}_1$ with a k-th order total variation penalty, P_k, for $k > 1$.

As discussed in Section 3.2, the main idea here is to approximate the indicator function, $I\{x > 0\}$, by what we will call the *k-th order soft indicator function*:

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

where b_{k-1} denotes the cardinal b-spline of order $k-1$, scaled to have support on $[-1, 1]$. b_{k-1} is a piecewise polynomial of order $k-1$, that is non-negative, and integrates to 1 [24]. Because of this, $I_k^\delta \in \mathcal{F}_k$; I_k^δ is monotonic with support on $[-\delta, \delta]$; and we have $I_k^\delta(-\delta) = 0$ and $I_k^\delta(\delta) = 1$.

Before we continue, we note that for the class $\mathcal{F}_k$ with our penalty P_k, we get an entropy as in (24) with $\alpha = 1/k$ [1]. Thus, the term depending on the entropy of our class (27) becomes

$$n^{\frac{-2}{2+\alpha}} P^{\frac{2\alpha}{2+\alpha}}(f) = n^{-2k/(2k+1)} P^{2/(2k+1)}(f). \tag{S3.1}$$

We will prove Lemma 3.1 first for piecewise constant functions with a single knot; then with multiple knots; and finally for general functions in $\mathcal{F}_1$.

S3.2 Estimating Piecewise Constant Functions With A Single Knot

First we consider estimating f^*, a piecewise constant function with a single jump. Without loss of generality suppose $f^*(x) = \beta_0 * I\{x > 0\}$. We use our *k-th order soft indicator function* to give approximating functions in $\mathcal{F}_k$: In particular, we choose $f_\delta^O \equiv \beta_0 I_k^\delta$.

It is straightforward to show that

$$\left\| f^* - f_\delta^O \right\|_n^2 \leq 2\beta_0^2\delta \qquad \text{and} \qquad P_k\left(f_\delta^O\right) \leq \frac{C\beta_0}{\delta^{k-1}}$$

for a fixed C. The first inequality follows because $f_\delta^O(x)$ is identical to f^* outside of $[-\delta, \delta]$, and monotonically moves from 0, to β_0, in that interval. The second follows from basic calculus (given in detail in Section S1.6)

Remembering our earlier entropy bound (S1.6), and recalling the result of Theorem 2.2, we now need to balance

$$\beta_0^2\delta(n) \qquad \text{and} \qquad n^{-2k/(2k+1)}\left[\delta(n)\right]^{-2(k-1)/(2k+1)}\beta_0^{2/(2k+1)}.$$

These terms are balanced by $\delta(n) = n^{-2k/(4k-1)}\beta_0^{-4k/(4k-1)}$. Plugging this in to (25) gives

$$\left\| \hat{f} - f^* \right\|_n^2 \leq \left\| \hat{f} - f_{\delta(n)}^O \right\|_n^2 + O_p\left(\lambda_n P\left(f_{\delta(n)}^O\right)\right) = O_p\left(n^{\frac{-2k}{4k-1}}\beta_0^{\frac{4k-2}{4k-1}}\right).$$

Noting that $\beta_0 = P_1(f^*)$, we have the rate in (38).

S3.3 Estimating a Piecewise Constant Function With Multiple Knots

We now generalize the result of the previous section to a function f^* with multiple jumps: $f^*(x) = \beta_0 + \sum_{j=1}^J \beta_j * I\left\{x > d_j\right\}$. We can approximate each jump by a k-th order soft indicator function; and define our approximator f_δ^O as the sum of all of these functions:

$$f_\delta^O(x) \equiv \beta_0 + \sum_{j=1}^{J} \beta_j I_k^\delta (x - d_j).$$

We first note that $f_\delta^O \in \mathcal{F}_k$. By the triangle inequality,

$$P_k\left(f_\delta^O\right) \leq \sum_{j=1}^{J} \beta_j P_k\left(I_k^\delta (x - d_j)\right) \leq \frac{C}{\delta^{k-1}} \sum_{j=1}^{J} \beta_j = \frac{C}{\delta^{k-1}} P_1(f^*),$$

where $P_1(f^*) = \sum_{j=1}^{J} \beta_j$ is the total variation of f^*. In addition, we have

$$\begin{aligned}
\left\|f^* - f_\delta^O\right\|_n &\leq \left\| \sum_{j=1}^{J} \beta_j * I\{x > d_j\} - \sum_{j=1}^{J} \beta_j I_k^\delta (x - d_j) \right\|_n \\
&\leq \sum_{j=1}^{J} \beta_j \left\| I\{x > d_j\} - I_k^\delta (x - d_j) \right\|_n \\
&\approx \sum_{j=1}^{J} \beta_j \sqrt{2\delta} \\
&= \sqrt{2\delta} P_1(f^*).
\end{aligned}$$

Thus, that $\left\|f^* - f_\delta^O\right\|_n^2 \lesssim \delta P_1(f^*)^2$. This exactly mirrors what we saw in the previous section. So choosing $\delta(n) = n^{-2k/(4k-1)} P_1(f^*)^{-4k/(4k-1)}$, again gives us the rate in (38).

One noteworthy aspect of the above result is that the number of knots does not show up in the rate — only the total variation shows up. This will be key in the next section, where we get identical bounds for general functions in $\mathcal{F}_1$.

S3.4 Estimating a General Function in $\mathcal{F}_1$

We now prove Lemma 3.1 in its general form. Suppose that f^* is any function in $\mathcal{F}_1$. Here we use the result of Birman and Solomyak [4] that for any δ, there exists a piecewise constant function function $\tilde{f}^\delta$ such that

$$\left\|f^* - \tilde{f}^\delta\right\|_n^2 \leq \delta P_1\left(f^*\right)^2 \qquad \text{and} \qquad P_1\left(\tilde{f}^\delta\right) \leq \tilde{C} P_1\left(f^*\right) \tag{S3.2}$$

for a constant $\tilde{C}$ that does not depend on f^*. More explicitly, $\tilde{f}^\delta(x) = \beta_{0,\delta} + \sum_{j=1}^{J(\delta)} \beta_{j,\delta} I\left\{x > d_{j,\delta}\right\}$, for some knots $d_{j,\delta}$ and heights $\beta_{j,\delta}$ that depend on δ (and f^*). Now, we use the same construction for f_δ^O as in Section S1.3.3, only with $\tilde{f}^\delta$ taking the place of f^*, i.e.,

$$f_\delta^O(x) \equiv \beta_{0,\delta} + \sum_{j=1}^{J(\delta)} \beta_{j,\delta} I_k^\delta\left(x - d_{j,\delta}\right).$$

From here we see that

$$P_k\left(f_\delta^O\right) \leq \frac{C}{\delta^{k-1}} P_1\left(\tilde{f}^\delta\right) \leq \frac{C_1}{\delta^{k-1}} P_1\left(f^*\right)$$

for some constant C_1, and,

$$\begin{aligned}
\left\|f^* - f_\delta^O\right\|_n &\leq \left\|f^* - \tilde{f}^\delta\right\|_n + \left\|\tilde{f}^\delta - f_\delta^O\right\|_n \\
&\leq= \sqrt{\delta} P_1\left(f^*\right) + \sqrt{2\delta} P_1\left(\tilde{f}^\delta\right) \\
&\leq \left(1 + \sqrt{2}\right) \sqrt{\delta} P_1\left(f^*\right).
\end{aligned}$$

Thus, using the same argument as before, we get the rate in (38).

S3.5 Estimating a function in $\mathcal{F}_{l+1}$ using P_k $(k > l+1 \geq 2)$

In this section, we prove Lemma 3.2 about the estimation of functions with $l+1$ order bounded variation, using P_k, where $k > l+1$; and $l \geq 1$. We will again prove this Lemma in stages: First for a spline with a single knot; then a spline with multiple knots; and finally an arbitrary element of $\mathcal{F}_{l+1}$.

S3.6 Estimating a Natural Spline of order l with 1 knot

Suppose $f^*(x) = \beta_0 x^l I(x \geq 0)$. Now, we approximate f^* by our representative $f_\delta^O(x) = \beta_0 \psi_{k,l}^\delta(x)$, with

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

as discussed in Section 3.3. Noting that $\frac{\partial^l}{\partial x^l}\psi_{k,l}^\delta(x) = l! I_k^\delta(x)$, and $\frac{\partial^l}{\partial x^l} x^l I(x \geq 0) = l! I(x \geq 0)$, this gives us that

$$\left|\frac{\partial^l}{\partial x^l}\left[f^* - f_\delta^O\right](x)\right| \leq \begin{cases} l!\beta_0 & x \in [-\delta, \delta] \\ 0 & x \notin [-\delta, \delta] \end{cases}$$

and that

$$P_k\left(f_\delta^O\right) = \int \left|\left(\frac{\partial^l}{\partial x^l} f_\delta^O\right)^{(k-l)}(x)\right| dx = \frac{C_1 \beta_0}{\delta^{k-l-1}},$$

for some constant C_1, which can again be seen from the discussion in Section S1.6. Note that here we use weak derivatives.

Now using repeated integration (l times), and the fact that $f^*(-\delta) = f_\delta^O(-\delta)$, we get that, for any x,

$$\begin{aligned}
\left|f^*(x) - f_\delta^O(x)\right| &= \left|\underbrace{\int_{-\delta}^{x} \cdots \int_{-\delta}^{t_2}}_{l \text{ times}} \frac{\partial^l}{\partial x^l}\left[f^* - f_\delta^O\right](t_1)\, dt_1 \cdots dt_l\right| \\
&\leq \int_{-\delta}^{x} \cdots \int_{-\delta}^{t_2} \left|\frac{\partial^l}{\partial x^l}\left[f^* - f_\delta^O\right](t_1)\right| dt_1 \cdots dt_l \\
&\leq \int_{-\delta}^{x} \cdots \int_{-\delta}^{t_2} \beta_0 l! I\left(-\delta \leq t_1 \leq \delta\right) dt_1 \cdots dt_l \\
&\leq C_2 \beta_0 \delta.
\end{aligned}$$

For some constant C_2. Thus we have that $\left\|f^*(x) - f_\delta^O(x)\right\|_n^2 \leq C_2 \beta_0^2 \delta^2$. This implies that we need to balance

$$\beta_0^2 \delta(n)^2 \qquad \text{and} \qquad n^{-2k/(2k+1)}\left[\delta(n)\right]^{-2(k-l-1)/(2k+1)} \beta_0^{2/(2k+1)},$$

which are balanced by $\delta(n) \sim n^{-\frac{k}{3k-l}} \beta_0^{-\frac{2k}{3k-l}}$. Plugging this in to (25) gives us our rate in (39).

S3.7 Estimating A Spline of Order l with multiple knots

Now suppose $f^*(x) = f_0^*(x) + \sum_{j=1}^{J} \beta_j \left(x - d_j\right)_+^l$, where f_0^* is an order l polynomial. Note that $P_{l+1}(f^*) = \sum_{j=1}^{J} \beta_j$. We will use the same method of construction/proof as in Section S1.3.3. We let f_δ^O be given by

$$f_\delta^O(x) \equiv f_0^*(x) + \sum_{j=1}^{J} \beta_j \psi_{k,l}^\delta\left(x - d_j\right).$$

Since P_k is a semi-norm, it obeys the triangle inequality; so,

$$P_k\left(f_\delta^O\right) \leq \sum_{j=1}^J \beta_j P_k\left(\psi_{k,l}^\delta\right) \leq \sum_{j=1}^J \beta_j \left(\frac{C_1}{\delta^{k-l-1}}\right) = \frac{C_1 P_{l+1}\left(f^*\right)}{\delta^{k-l-1}}.$$

Additionally, using the arguments of Section S1.3.6, we have

$$\begin{aligned}\left\|f^* - f_\delta^O\right\|_n &\leq \sum_{j=1}^J \left\|\beta_j\left(x-d_j\right)_+^l - \beta_j \psi_{k,l}^\delta\left(x-d_j\right)\right\|_n \\ &\leq \sum_{j=1}^J \beta_j \left\|\left(x-d_j\right)_+^l - \psi_{k,l}^\delta\left(x-d_j\right)\right\|_n \\ &\leq \sum_{j=1}^J \beta_j C_2 \delta = C_2 \delta P_{l+1}\left(f^*\right).\end{aligned}$$

Thus, we have $\left\|f^* - f_\delta^O\right\|_n^2 \leq C_2^2 \delta^2 P_{l+1}\left(f^*\right)^2$. Using the same calculation and choice of δ as in the previous section we get the rate in (39).

S3.8 Estimating a general function in $\mathcal{F}_{l+1}$

We now prove Lemma 3.2 in its general form. Suppose that f^* lives in $\mathcal{F}_{l+1}$, the class of bounded $l+1$-th order total variation. This is equivalent to saying that $f_l^*(x) = \frac{\partial^l}{\partial x^l} f^*(x)$ is in $\mathcal{F}_1$. Using the result of Birman and Solomyak [4], for any $\delta > 0$ there exists a piecewise constant function $\tilde{f}_l^\delta$ such that

$$\left\|f_l^* - \tilde{f}_l^\delta\right\|_\infty \leq \delta P_{l+1}\left(f^*\right) \qquad \text{and} \qquad P_1\left(\tilde{f}_l^\delta\right) \leq \tilde{C} P_1\left(f_l^*\right) = \tilde{C} P_{l+1}\left(f^*\right) \tag{S3.3}$$

As before, we can explicitly write $\tilde{f}_l^\delta = \sum_{j=1}^{J(\delta)} l!\,\beta_{j,\delta} I\left\{x > d_{j,\delta}\right\}$ for some knots $d_{j,\delta}$ and heights $\beta_{j,\delta}$ that depend on δ (and f^*). Note that we include an $l!$ term in the representation.

From here we define $\tilde{f}^\delta$ by

$$\tilde{f}^\delta(x) \equiv f_0^*(x) + \sum_{j=1}^{J(\delta)} \beta_{j,\delta}\left(x - d_{j,\delta}\right)_+^l,$$

where $f_0^*(x)$ is an l-th order polynomial whose derivatives up to order l (including order 0) agree with f^* at $x = -1$. We note that $\frac{\partial^l}{\partial x^l}\tilde{f}^\delta(x) = \tilde{f}_l^\delta(x)$. In addition, using simple integration, as in Section S1.3.6, we can show that $\left\|f^* - \tilde{f}^\delta\right\|_n \leq C_1 \delta P_{l+1}\left(f^*\right)$, for some C_1.

We define our representative as

$$f_\delta^O(x) \equiv f_0^*(x) + \sum_{j=1}^{J(\delta)} \beta_{j,\delta}\psi_{k,l}^\delta\left(x - d_{j,\delta}\right)$$

Using the same argument as in S1.3.7 we see that

$$P_k\left(f_\delta^O\right) \leq \frac{C_2 P_{l+1}\left(f^*\right)}{\delta^{k-l-1}}$$

and

$$\begin{aligned}
\left\|f^* - f_\delta^O\right\|_n &\leq \left\|f^* - \tilde{f}_\delta\right\|_n + \left\|\tilde{f}_\delta - f_\delta^O\right\|_n \\
&\leq C_1 \delta P_{l+1}\left(f^*\right) + C_2 \delta P_{l+1}\left(f^*\right) \\
&= C_3 \delta P_{l+1}\left(f^*\right).
\end{aligned}$$

This mirrors the result from Section S1.3.6. Thus for the same choice of $\delta(n)$ we get the rate in (39). This can be extended estimating general $f^* \in \mathcal{F}_{l+1}$ using essentially identical arguments as in Sections S1.3.8 and S1.3.7.

S4. Details for Estimation with Sobolev Penalties

We now sketch similar results when we use Sobolev Penalties in our estimation procedure and/or when the true function lies in a class of bounded first order Total Variation.

First we consider estimating f^*, a piecewise constant function with a single jump using $P(\cdot) = P_k^d$ for $d > 1, k \geq 1$. Without loss of generality suppose $f^*(x) = \beta_0 * I\{x > 0\}$. We now use our $k+1$-*th order soft indicator function* to give approximating functions in $\mathcal{F}_k^d$: In particular, we choose $f_\delta^O \equiv \beta_0 I_{k+1}^\delta$. Note in the case of a total-variation penalty ($d = 1$) we were able to use a k-th order soft indicator (and got a correspondingly better rate)

As before, it is straightforward to show that

$$\left\| f^* - f_\delta^O \right\|_n^2 \leq 2\beta_0^2 \delta \qquad \text{and} \qquad P_k^d\left(f_\delta^O\right) \leq \frac{C\beta_0}{\delta^k}$$

The second inequality follows again from basic calculus (given in detail in Section S1.6)

The entropy of the k-th order sobolev class is also given by (S1.6) [25], and recalling the result of Theorem 2.2, we now need to balance

$$\beta_0^2 \delta(n) \qquad \text{and} \qquad n^{-2k/(2k+1)} \left[\delta(n)\right]^{-2k/(2k+1)} \beta_0^{2/(2k+1)}.$$

These terms are balanced by $\delta(n) = n^{-2k/(4k+1)} \beta_0^{-4k/(4k+1)}$. Plugging this in to (25) gives

$$\left\| \hat{f} - f^* \right\|_n^2 \leq \left\| \hat{f} - f^O_{\delta(n)} \right\|_n^2 + O_p\left(\lambda_n P\left(f^O_{\delta(n)}\right)\right) = O_p\left(n^{\frac{-2k}{4k+1}} \beta_0^{\frac{4k+2}{4k+1}}\right).$$

Noting that $\beta_0 = P_1(f^*)$, we have the rate in (38).

To extend this to estimating a general function in $f^* \in \mathcal{F}_1$, we first extend the above result to the estimation of piecewise constant functions with multiple knots. The proof follows almost exactly as the proof in Section S1.3.3. Here again, we can employ the triangle inequality because P_k^d is a norm. Finally, by mirroring the argument in Section S1.3.4 we get the result in Lemma 4.3.

S5. Estimating a function in $\mathcal{F}_{l+1}$ with P_k^d

We now bound our convergence rates when using a kth-order Sobolev penalty, where the true function lies in a class of bounded $l+1$-th order total variation for $2 \leq l+1 < k$.

We begin, as in Section S1.3.6, by restricting f^* to be a Natural Spline of order l with 1 knot. Suppose $f^*(x) = \beta_0 x^l I(x \geq 0)$. Now, we approximate f^* by our representative $f_\delta^O(x) = \beta_0 \psi_{k+1,l}^\delta(x) \in \mathcal{F}_k^d$, with

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

We note that in Section S1.3.6 we were able to use $\psi_{k,l}^\delta$; however $\psi_{k,l}^\delta \notin \mathcal{F}_k^d$. As in Section S1.3.6, we get that

$$\left\| f^*(x) - f_\delta^O(x) \right\|_n^2 \leq C_1 \beta_0^2 \delta^2 \qquad \text{and} \qquad P_k\left(f_\delta^O\right) \leq \frac{C_2 \beta_0}{\delta^{k-l}}$$

for some constants C_1, C_2. This implies that we need to balance

$$\beta_0^2 \delta(n)^2 \qquad \text{and} \qquad n^{-2k/(2k+1)} \left[\delta(n)\right]^{-2(k-l)/(2k+1)} \beta_0^{2/(2k+1)},$$

which are balanced by $\delta(n) \sim n^{-\frac{k}{3k-l+1}} \beta_0^{-\frac{2k+1}{3k-l+1}}$. Plugging this in to (25) gives us

$$\left\| \hat{f} - f^* \right\|_n^2 \leq \left\| \hat{f} - f_{\delta(n)}^O \right\|_n^2 + O_p\left(\lambda_n P\left(f_{\delta(n)}^O\right)\right) = O_p\left(n^{\frac{-2k}{3k-l+1}} \beta_0^{\frac{2k-2l}{3k-l+1}}\right).$$

This is the rate we have in Lemma 4.4. Mirroring the arguments of Sections S1.3.8, and S1.3.7, we can extend this to estimating arbitrary functions, f^*, in $\mathcal{F}_{l+1}$.

S6. Properties of our B-spline representative

Here we discuss some properties of B-splines that were used in constructing our estimates. We first let b_{k-1} denote the cardinal b-spline of order $k-1$, scaled to have support on $[-1, 1]$. b_{k-1} is a piecewise $k-1$ order polynomial, that is non-negative, and integrates to 1 [24].

Before moving further, for $m < k-1$, let $b_{k-1}^{(m)}(x_0)$ denote

$$b_{k-1}^{(m)}(x_0) \equiv \frac{\partial^m}{\partial x^m} b_{k-1}(x)\Big|_{x=x_0}$$

and let $H_{k-1,m-1} = \int \left| b_{k-1}^{(m-1)}(x) \right| dx$ for $m \leq k$, where this is defined based on weak derivatives for $k = m$.

Now we will consider properties of $f^\delta(x) \equiv \beta_0 * \delta^{-1} \int_{-\infty}^{x} b_{k-1}\left(\frac{t}{\delta}\right) dt$. We note that f^δ is a k-th order spline (only its last derivative changes non-smoothly); and we have $f^\delta(x) = 0$ for $x \leq -\delta$ and $f^\delta(x) = \beta_0$ for $x \geq \delta$ (by properties of b_{k-1}).

We also note that

$$\begin{aligned}
\frac{\partial^m}{\partial x^m} f^\delta(x)\Big|_{x=x_0} &= \beta_0 \delta^{-1} \left(\frac{\partial^{m-1}}{\partial x^{m-1}} b_{k-1}\left(x/\delta\right)\Big|_{x=x_0} \right) \\
&= \beta_0 \delta^{-1} \left(\frac{1}{\delta^{m-1}} \right) b_{k-1}^{(m-1)}\left(x_0/\delta\right) \\
&= \left(\frac{\beta_0}{\delta^m} \right) b_{k-1}^{(m-1)}\left(x_0/\delta\right)
\end{aligned}$$

. Thus, for $m \leq k$ we have

$$\begin{aligned}\int \left|\frac{\partial^m}{\partial x^m} f^\delta(x)\right| dx &= \frac{\beta_0}{\delta^m} \int \left|b_{k-1}^{(m-1)}(x/\delta)\right| dx \\ &= \frac{\beta_0}{\delta^{m-1}} \int \left|b_{k-1}^{(m-1)}(x)\right| dx \\ &= \frac{(H_{k-1,m-1})\,\beta_0}{\delta^{m-1}}\end{aligned}$$

where this is defined based on weak derivatives for $m = k$.

Also, note that for $m < k$, and any $d > 1$ an identical argument can be used to show

$$\left\{\int \left|\frac{\partial^m}{\partial x^m} f^\delta(x)\right|^d dx\right\}^{1/d} = \frac{\left(H_{k-1,m-1}^d\right)\beta_0}{\delta^{m-1}}$$

where we define $H_{k-1,m-1}^d = \left\{\int \left|b_{k-1}^{(m-1)}(x)\right|^d dx\right\}^{1/d}$. Here we need $m < k$, because the integral diverges to ∞ using weak derivatives for $m = k$.

S7. Bounds for Empirically Selected λ

Here we extend the discussion of bounds for the penalized estimator with λ selected empirically, that began in Section 3.1.

To begin, we consider why the optimal λ should be a function of both k (the smoothness induced by our penalty) and l (the true underlying smoothness of f^*). We build our intuition from a simpler scenario: Kernel Density Estimation (KDE) in $\mathbb{R}^1$

Suppose, we use a k-th order kernel to estimate a density from iid observations. Imagine that the true density g^* has only $l < k$ bounded derivatives. In this case, our KDE can give a minimax optimal estimate (over the class of densities with bounded derivatives of order l). However, to do this, we must use a bandwidth that depends on l. This is because, for a given bandwidth, the variance of our estimator will be the same, regardless of l; but the bias will be a function l (smoother g^* induce less bias at a given bandwidth). Thus, to balance bias and variance we must choose lower bias/higher variance estimates for less smooth functions.

Now, let us relate this back to the current problem. For penalized estimators, λ determines the bias/variance tradeoff of an estimator (lower λ indicates a lower bias, higher variance estimate). In this case, if l is smaller, that would imply that f^* is less smooth, and thus we need a smaller λ-value. This can also be directly observed in Theorem 2.2, where λ_n is selected to be proportional to $(P(f_n^O))^{-\left(\frac{2-\alpha}{2+\alpha}\right)}$: As $P(f_n^O)$ increases, we need λ_n to decrease (if the function is more rough, we shouldn't penalize roughness as much). In addition, our approximation theory results indicate that as l gets smaller, it takes a function f_n^O with larger $P_k(f_n^O)$ to approximate f^* well.

Even though the indicated λ depends on the unknown quantity l, these *oracle bounds* can still be useful in proving bounds for estimators with

λ selected by split-sample validation. In particular, suppose our data is partitioned into a training subset, and a validation subset. For any given λ, $\hat{f}^{\lambda}$ is calculated by minimizing (12) (with that given λ) over the training data. λ_V is then selected as $\mathrm{argmin}_{\lambda\in\Lambda}\left\|\hat{f}^{\lambda}-y\right\|_{n,V}$, the minimizer of the empirical error over the validation data; where Λ is a search space for λ. Using recent work [11], one can shown that

$$\left\|\hat{f}^{\lambda_V}-f^*\right\|_{n,V}^2 \leq \min_{\lambda\in\Lambda}\left\|\hat{f}^{\lambda}-f^*\right\|_{n,V}^2 + R_n(\Lambda) \tag{S7.1}$$

where $R_n(\Lambda)$ is some excess error that depends on the complexity of Λ. Thus, if $\Lambda \equiv [\lambda_{min}, \lambda_{max}]$ with λ_{min} shrinking sufficiently quickly to 0, then $\min_{\lambda\in\Lambda}\left\|\hat{f}^{\lambda}-f^*\right\|_{n,V}^2$ is upper-bounded (we believe in some cases, sharply) by the results in Lemma 3.1 and Lemma 3.2. Characterizing the behaviour of $R_n(\Lambda)$ here would result in upper bounds on the error of the estimator obtained by solving the penalized regression problem (12) with λ chosen by split-sample validation. In particular [11] show that if the penalty is a squared-sobolev-seminorm, and if λ_{min} decreases at a polynomial rate, then $R_n(\Lambda)$ is negligible. With slight modification (to move to the sobolev semi-norm), this could be used to show that with Sobolev semi-norm penalties, using split sample validation to select λ would result in an estimator that achieves our oracle rate. In this manuscript, we focus on bounding the oracle error — we leave engaging further with error for empirically selected

λ, eg. using (S1.9), to future work.

S8. Additional Simulations

Here, we extend the simulation settings of Section 5 in two ways: First we use non-gaussian errors. In particular, we use errors that are uniformly distributed; and double-exponential. In both cases we center/scale our errors to have mean 0 and variance 1. Our second modification is to include additional functions for f^*. In particular, here we use still use a piecewise constant and linear function, but now generate those functions to have knots at a several (5 and 15 in our scenarios) random uniformly-generated locations (with random-sized jumps, also uniformly generated): These functions are given below in Figures 5, and 7.

We see the results in the Figures 4, 6, and 8. We note that, for the piecewise constant and linear functions with a single knot, when we generate data with non-gaussian errors, the results remain largely unchanged, as seen in Figure 4. The multi-knot functions also exhibit similar behaviour as seen in Figures 6 and 8; with best performance for penalties that match the maximal smoothness of the function (P_1 for the piecewise constant and P_2 for the piecewise linear), but still reasonable performance (and prediction consistency) when overly ambitious penalties are employed. In particular using P_3 for the piecewise-linear function gives quite strong performance. It is also worth noting that these results are remarkably consistent across the 3 error distributions.

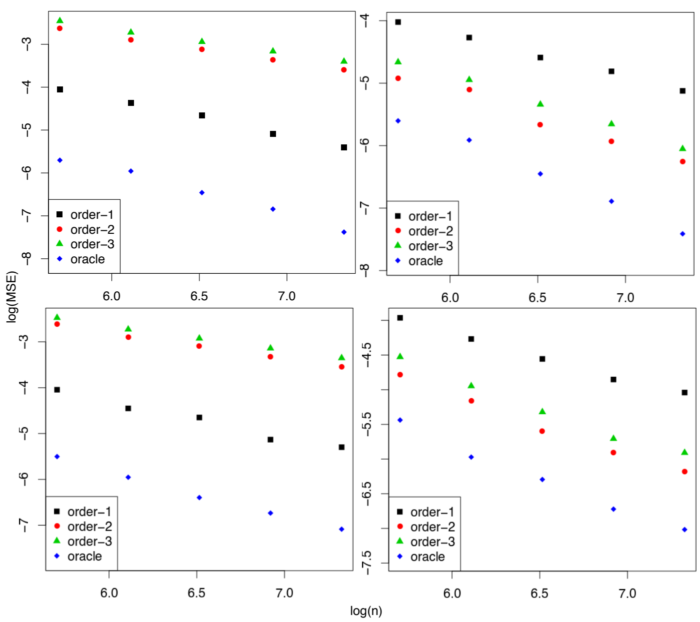

Figure 4: Average log(MSE) vs. $\log(n)$ for estimators with total variation penalties of degree 1, 2 and 3, along with a parametric oracle. In the left panels, data were generated using the regression function $f^*(x) = 3 * I(x > 0.5)$; in the right panel, $f^*(x) = 3(x - 0.5)_+$ was used. In the top panel, ϵ_i were uniformly distributed; in the bottom, from a double-exponential distribution. MSE was calculated as the average over 100 simulations for each $n_j = 200 * 1.5^j$ for $j = 1, \ldots, 5$.

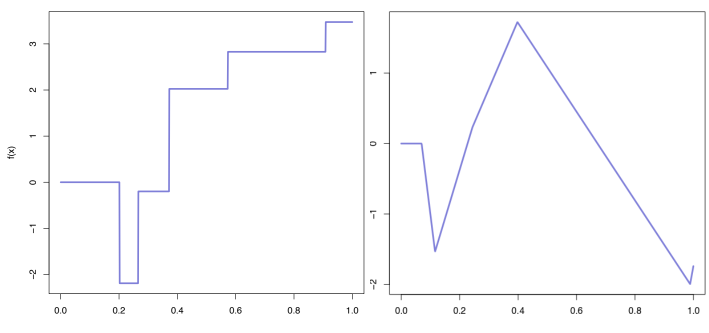

Figure 5: Two additional f^* functions used in simulations. On the left we have a piecewise constant function (with 5 knots); on the right, we have a piecewise linear function (also with 5 knots).

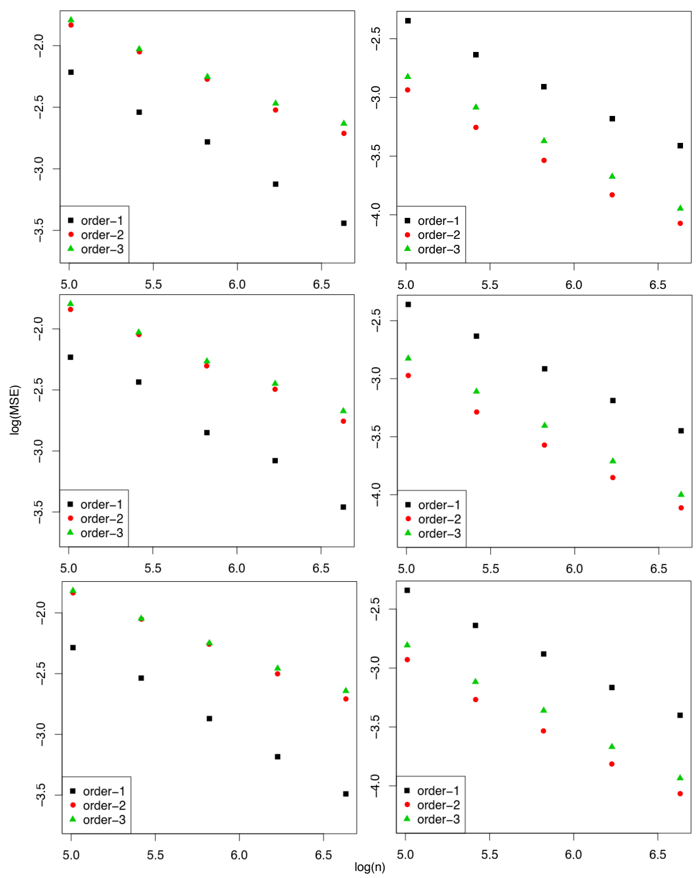

Figure 6: Average log(MSE) vs. $\log(n)$ for estimators with total variation penalties of degree 1, 2 and 3 estimating piecewise polynomial functions with 5 knots. In the left panels, data were generated using the piecewise constant regression function seen in the left panel of Figure 5; in the right panel, the piecewise linear function in the right panel of Figure 5 was used. In the top panel, ϵ_i were uniformly distributed; in the middle, from a double-exponential distribution, and in the bottom, from a gaussian. MSE was calculated as the average over 100 simulations for each $n_j = 100 * 1.5^j$ for $j = 1, \ldots, 5$.

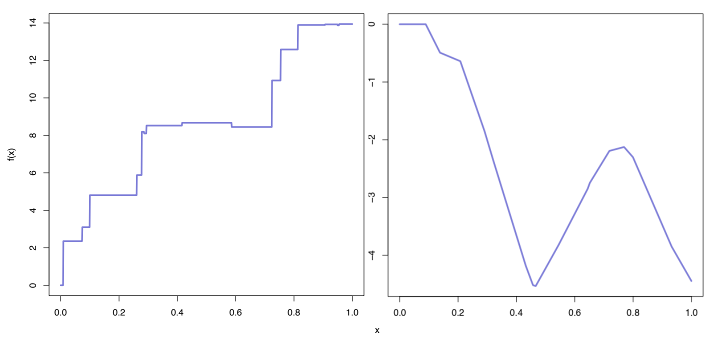

Figure 7: Two additional f^* functions used in simulations. On the left we have a piecewise constant function (with 15 knots); on the right, we have a piecewise linear function (also with 15 knots).

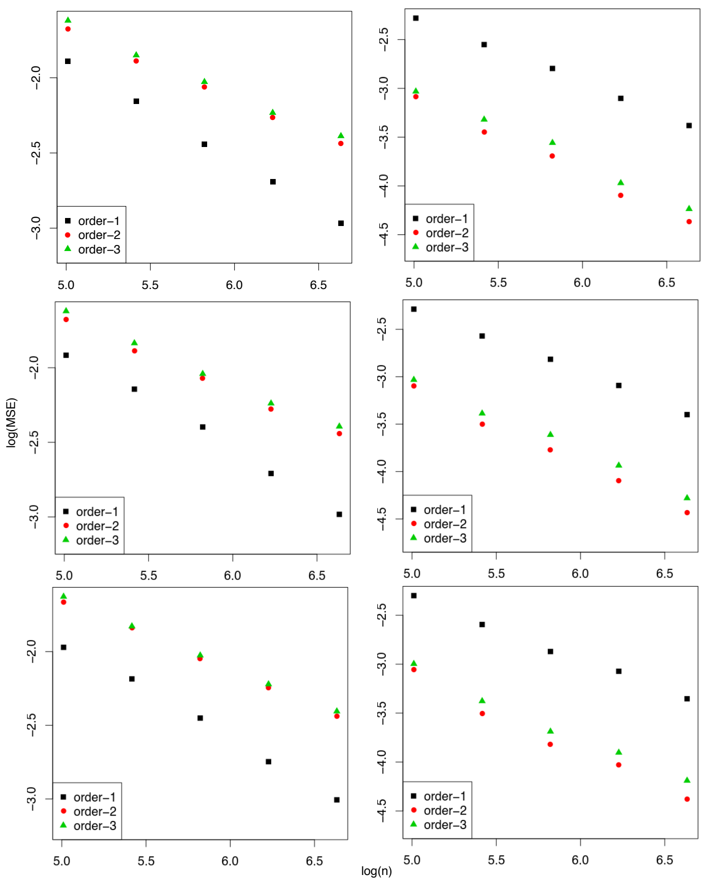

Figure 8: Average log(MSE) vs. log(n) for estimators with total variation penalties of degree 1, 2 and 3 estimating piecewise polynomial functions with 15 knots. In the left panels, data were generated using the piecewise constant regression function seen in the left panel of Figure 7; in the right panel, the piecewise linear function in the right panel of Figure 7 was used. In the top panel, ϵ_i were uniformly distributed; in the middle, from a double-exponential distribution, and in the bottom, from a gaussian. MSE was calculated as the average over 100 simulations for each $n_j = 100 * 1.5^j$ for $j = 1, \ldots, 5$.