Smooth Regression

January 30, 2025

In this homework, we'll briefly review Bounded Variation Regression and then explore Lipschitz Regression, another form of smooth regression. We will focus on the one-dimensional case, although it extends very naturally to higher dimensions. Then we'll look into rates of convergence, comparing this new method to the stuff we've been using.

1 Review of Bounded Variation Regression

In class, we talked about using least squares regression to fit a function of *bounded total variation*. If we are fitting $\mu(x) = \mathrm{E}[Y_i \mid X_i = x]$ for covariates $X_i \in [0, 1]$, this estimator is

$$\begin{aligned} \hat{\mu} = & \operatorname*{argmin}_{\substack{m \\ \rho_{TV}(m) \le B}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - m(X_i)\}^2 && \text{where} \\ \rho_{TV}(m) = & \int_0^1 |m'(x)| dx && \text{for differentiable } m \\ = & \sup_{\substack{\text{increasing sequences} \\ x_1 \le x_2 \le \ldots \le x_k \\ x_1 \ldots x_k \in [0,1]}} \sum_j |m(x_{j+1}) - m(x_j)| && \text{generally .} \end{aligned} \tag{1}$$

The set of functions we're optimizing over, those with $\rho_{TV}(m) \le B$, is a set of functions that doesn't vary too much in total. It does, however, include both functions that vary slowly throughout the interval $[0, 1]$ and those that vary quickly for a small part of it.

Exercise 1 *To get a sense of what the constraint $\rho_{TV}(m) \le B$ means, calculate $\rho_{TV}(m)$ for the following functions on $[0, 1]$. These are repeats from class.*

1. $m(x) = x$
2. $m(x) = x^2$
3. $m(x) = e^x$
4. $m(x) = \sin(\pi x)$

5. $m(x) = \begin{cases} 0 & \text{if } x < 1 \\ 1 & \text{if } x = 1 \end{cases}$

6. $m(x) = \sin(1/x)$

Exercise 2 *For the following, which are a bit subtler, give an upper bound on $\rho_{TV}(m)$. If it's infinite, explain why.*

1. $m(x) = x \sin(1/x)$

2. $m(x) = x^2 \sin(1/x)$

3. $m(x) = x^{3/2} \sin(1/x)$

Hint: It might be hard to find the upper bound by looking at the graph of some of these functions. Instead, find the derivative and a corresponding upper bound for it, if possible. Your upper bound doesn't have to be tight. When you are bounding a sum, use the triangle inequality by adding the upper bound for each term.

2 Lipschitz Regression

In some cases, it may be implausible that $\mu(x)$ varies quickly anywhere. In that case, we may prefer to fit a *Lipschitz function*, for example by solving the following least squares problem.

$$\begin{aligned} \hat{\mu} &= \operatorname*{argmin}_{\substack{m \\ \rho_{Lip}(m) \leq B}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - m(X_i)\}^2 \quad \text{where} \\ \rho_{Lip}(m) &= \sup_{x \in [0,1]} |m'(x)| \qquad \text{for differentiable } m \\ &= \sup_{\substack{x_1, x_2 \in [0,1] \\ x_1 \neq x_2}} \frac{|m(x_2) - m(x_1)|}{|x_2 - x_1|} \qquad \text{generally .} \end{aligned} \tag{2}$$

We call $\rho_{Lip}(m)$ the *Lipschitz constant* of the function m. Let's interpret the general definition visually. It's a maximum of the absolute value of the slope of the functions's secants.

Equivalence. Our definition for differentiable functions is equivalent because (i) derivatives are included in the set of slopes we're maximizing over, as they are the slopes of tangents, which are just very short secants (ii) every slope in this set is equal to a derivative, as the mean value theorem tells us that the slope of the secant drawn from $x = a$ to $x = b$ is equal to the derivative of the function at some point between a and b.

2.1 Finding Lipschitz Constants

Exercise 3 *To get a sense of what this new type of constraint $\rho_{Lip}(m) \leq B$ means, calculate $\rho_{Lip}(m)$ for the examples from Exercise 1. Bound it or explain why it's infinite for the examples from Exercise 2. Is $\rho_{TV}(m) \leq \rho_{Lip}(m)$ for all of these examples? If so, either prove that it's true for all functions m on $[0,1]$ or find a counterexample.*

2.2 Fitting the Lipschitz Model

As usual, we'll start by solving for a function $\hat{\mu}_{|\mathcal{X}}$ on the sample $\mathcal{X} = \{X_1 \dots X_n\}$, then extend it to the real line. Just like in the bounded variation lab, we'll translate our seminorm ρ_{Lip} into a seminorm on functions $m : \mathcal{X} \to \mathbb{R}$ simply by replacing the 'for all ... $\in \mathbb{R}$' (or $[0,1]$) with a 'for all ... $\in \mathcal{X}$'. Like this.

$$\begin{aligned} \hat{\mu}_{|\mathcal{X}} &= \operatorname*{argmin}_{\substack{m \\ \rho_{Lip|\mathcal{X}}(m) \leq B}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - m(X_i)\}^2 \quad \text{where} \\ \rho_{Lip|\mathcal{X}}(m) &= \sup_{x \neq x' \in \mathcal{X}} \frac{|m(x) - m(x')|}{|x - x'|} \\ &= \sup_{\substack{\text{pairs } i,j \in 1 \dots n \\ \text{with } X_i \neq X_j}} \frac{|m(X_i) - m(X_j)|}{|X_i - X_j|}. \end{aligned} \tag{3}$$

This is something we can handle. This depends on the values of $m(x)$ only at the observed data points $x \in \{X_1 \dots X_n\}$, so we can implement it as an optimization over a vector $\vec{m} \in \mathbb{R}^n$ with the interpretation that $\vec{m}_i = m(X_i)$. The constraint $\rho_{Lip|\mathcal{X}}(m) \leq B$ can be expressed as a set of constraints on $\vec{m}_i - \vec{m}_j$ for pairs i, j.

Exercise 4 *Rewrite this problem as a constrained optimization over the vector $\vec{m}$. Try to do it so what you've written translates straightforwardly into CVXR code.*

Tip. There is a smaller set of constraints that implies the full set in (3). We'll get there in Exercise 10. For now, use the full set, like we did until the section 'Optional Exercise: Optimization' in the monotone regression lab.

Tip. If you want to keep things simple, go ahead and assume that $X_1 \dots X_n$ take on n distinct values, just like we did at the beginning of the monotone regression lab. If you want more generally applicable code, take a look at how we use `invert.unique` in the monotone regression lab to handle duplicate values.

Tip. CVXR seems to be having some trouble with this one if we use division in our constraint, so don't. To write your constraint without division, observe that the following set of constraints are equivalent: (i) $\max_{i \leq n} |u_i / v_i| \leq B$, (ii) $|u_i|/|v_i| \leq B$ for all $i \in 1 \dots n$, and (iii) $|u_i| \leq B|v_i|$ for all $i \in 1 \dots n$.

Exercise 5 *Implement that optimization in* **R**. *That is, write an* **R** *function* **lipreg** *analogous to* **monotonereg** *from the monotone regression lab that solves* (3). *Then, from the six distributions described below, sample* $n = 100$ *observations* $(X_1, Y_1) \dots (X_n, Y_n)$ *and use your code to calculate predictions* $\hat{\mu}(X_1) \dots \hat{\mu}(X_1)$ *based on the solution to* (3) *with variation bound* $B = 1$. *Each time, plot your predictions on top of the data, i.e., make a single scatter plot showing both your predictions* $(X_i, \hat{\mu}(X_i))$ *and your observations* (X_i, Y_i). *Turn in those six plots, labeling each with the signal used, as your solution to this exercise.*

We'll sample observations around six signals.

1. A step, $\mu(x) = 1(x >= .5)$.
2. A line, $\mu(x) = x$.
3. A vee, $\mu(x) = (x - .5)1(x \geq .5)$.
4. A sine, $\mu(x) = \sin(\pi x)$.
5. A damped rapidly oscillating curve, $\mu(x) = x \sin(1/x)$.
6. A more-damped rapidly-oscillating curve, $\mu(x) = x^{3/2} \sin(1/x)$.

For each, we'll work with independent and identically distributed observations $(X_1, Y_1) \dots (X_n, Y_n)$ where X_i is drawn from the uniform distribution on $[0, 1]$ and $Y_i = \mu(X_i) + \varepsilon_i$ for ε_i drawn from the normal distribution with mean zero and standard deviation $\sigma = 1/10$.

Exercise 6 *Revisit the curves* $\hat{\mu}$ *you fit in the last exercise. For each, answer these questions.*

1. *Does it fit the data?*
2. *If not, what — if anything — could we do to fit the data better?*

Then, if there is something you can do, do it and include the resulting plot.

2.3 Filling in the gaps

At this point, you have an estimator $\hat{\mu}_{|\mathcal{X}}$ that minimizes squared error among the functions m satisfying $\rho_{Lip|\mathcal{X}}(m) \leq B$. This lets us plot some isolated points. But we want a complete curve $\hat{\mu}(x)$ for $x \in [0, 1]$ that satisfies $\rho_{Lip}(\hat{\mu}) \leq B$, and we want it to be the best-fitting such curve, i.e., we want the solution to (2).

To do this, we'll use a *piecewise-linear extension* of $\hat{\mu}_{|\mathcal{X}}$. That is, having sorted X_i into increasing order, we will define $\hat{\mu}(x)$ everywhere on $[X_1, X_n]$ by drawing line segments between successive points $\{X_i, \hat{\mu}(X_i)\}$ and $\{X_{i+1}, \hat{\mu}(X_{i+1})\}$, and extend the leftmost and rightmost segment to fill the intervals $[0, X_1]$ and $[X_n, 1]$.[1] This gives us a piecewise-linear solution to (3). First, we'll implement it. Then we'll verify that it is, in fact, a solution to (2).

Exercise 7 *Briefly explain why piecewise-constant extension would not give us a solution to* (2)*. A sentence or a sketch should do.*

Tip. Think about Exercise 3.

2.3.1 Implementation

Exercise 8 *Write out a formula for the piecewise-linear curve $\hat{\mu}(x)$ in terms of $\hat{\mu}(X_1) \dots \hat{\mu}(X_n)$. Then implement it and add the curve $\hat{\mu}(x)$ for $x \in [0, 1]$ to your plots from the last exercise.*

Tip. For coding a piecewise linear function, try to modify the function `predict.piecewise.constant` from the bounded variation lab.

2.3.2 Verification

Exercise 9 *Consider any pair $x < x'$. Prove that for any piecewise-linear function m with breaks at $X_1 \dots X_n$, the secant slope $\{m(x') - m(x)\}/(x' - x)$ between these points is a* weighted average *of the slopes $\{m(X_{j+1}) - m(X_j)\}/(X_{j+1} - X_j)$ of the segments that lie between them. Briefly explain why this implies that our piecewise-linear solution $\hat{\mu}$ satisfies $\rho_{Lip}(\hat{\mu}) = \rho_{Lip|\mathcal{X}}(\hat{\mu}_{|\mathcal{X}})$ and why this implies that $\hat{\mu}$ solves* (2)*.*

Tip. Break the 'explain' part of this down into feasibility and optimality, like we did in the bounded variation regression lab.

2.4 Optimized Fitting

We can speed up our fitting code by simplifying our set of constraints by hand. In particular, I claim that you get the same solution if you impose the constraint $|m(X_i) - m(X_j)|/|X_i - X_j| \leq B$ for adjacent points. That is, if the points $X_1 \dots X_n$ are sorted in increasing order, it's equivalent to impose the constraint for the pairs $(i, j = i + 1)$.

Exercise 10 *Prove it! Then implement it and check that your solution agrees with the one you got before using the all-pairs constraint. Include the proof as your solution. No need to turn in code, but you'll want this faster implementation later.*

Tip. The proof should be easy. Use the weighted-average idea from the last exercise.

3 Rates of Convergence

Now we've got three nonparametric regression models: monotone curves, bounded variation curves, and lipschitz curves. To keep things simple, we'll be working with data sampled around one signal: $\mu(x) = x$. That is, we'll work with independent and identically distributed observations $(X_1, Y_1) \dots (X_n, Y_n)$ where X_i is drawn from uniform distribution on $[0, 1]$ and $Y_i = \mu(X_i) + \varepsilon_i$ for ε_i drawn independently from the normal distribution with mean zero and standard deviation $\sigma = .5$.

Tip. What we're doing here is taking what we did at the end of the convergence rates lab, simplifying it by using only one signal instead of four, and then adding two new regression models. Use the lab's solution as a starting point.

Exercise 11 *Draw a sample of size $N = 1600$ from this distribution. To get samples of sizes $n = \{25, 50, 100, 200, 400, 800, 1600\}$, use the first 25, 50, etc. observations.*

At all of these sample sizes, fit a line, an increasing curve, and bounded variation and lipschitz curves with budgets $B = 1$. Calculate sample MSE $\|\hat{\mu} - \mu\|^2_{L_2(Pn)}$ and population MSE $\|\hat{\mu} - \mu\|^2_{L_2(P)}$ for each. Repeat this ten times and average the results to get estimates of expected sample MSE and expected population MSE at each sample size n. Include plots of these as a function of n as your solution.

Tip. This can be slow for larger samples. Try it out for samples of size $25 \dots 400$ before adding in $n = 800$ and $n = 1600$.

Let's try to summarize these plots by rates of convergence.

Exercise 12 *For each of your four regression models, use `nls` to fit a curve of the form $m(n) = \alpha n^{-\beta}$ to $RMSE = \sqrt{MSE}$ where MSE is your estimate of expected population mean squared error from the last exercise. Repeat for expected sample mean squared error.*

Plot the resulting predictions of MSE, $\hat{m}(n)^2$, on top of your actual MSE curves from the the previous exercise to check their accuracy. Include these plots and report these rates of convergence $\hat{\beta}$ as your solution. Briefly comment on what you see, too.

Notes

[1]I'm going to stop writing $\hat{\mu}_{|\mathcal{X}}$ all the time from here on. Since we're talking about an extension, we know that $\hat{\mu}(X_i) = \hat{\mu}_{|\mathcal{X}}(X_i)$, so I'll write that to reduce notational clutter.