Week 14 Homework: Covering Numbers

QTM 490R: Machine Learning Theory

1 Introduction

This week, we'll bound the gaussian width of neighborhoods in a bounded variation regression model.

$$\begin{aligned} \mathcal{M} &= \{m \ : \ m(x) \in [0,1] \ \text{ and } \ \rho_{TV}(m) \leq 1\} \\ &\text{where} \quad \rho_{TV}(m) = \sup_{\substack{\text{finite sequences} \\ 0=x_0 \leq \ldots \leq x_K = 1}} \sum_{j=1}^{K} |x_j - x_{j-1}|. \end{aligned} \tag{1}$$

To do this, we'll use *Dudley's Integral Bound*.

$$\mathrm{w}(\mathcal{V}) \lesssim \int_0^\infty \sqrt{\log(K_\epsilon)} d\epsilon \quad \text{if } \mathcal{V} \text{ has an } \epsilon\text{-cover containing } K_\epsilon \text{ curves.} \tag{2}$$

That means what we've got to do is find an ϵ-cover for a neighborhood $\mathcal{M}_s$ in this model. In fact, we'll do something a bit simpler. We'll find an ϵ-cover for the whole model and observe that (i) this is an ϵ-cover for neighborhoods $\mathcal{M}_s$ of any radius s and (ii) if $\epsilon \geq s$ then $\mathcal{M}_s$ has an ϵ-cover of size $K_\epsilon = 1$. Because $\log(1) = 0$, this implies a version of Dudley's Integral Bound with K_ϵ corresponding to a cover of $\mathcal{M}$ itself and $\epsilon = s$ as an upper limit of integration.

$$\mathrm{w}(\mathcal{M}_s) \lesssim \int_0^s \sqrt{\log(K_\epsilon)} \quad \text{if } \mathcal{M} \text{ has an } \epsilon\text{-cover containing } K_\epsilon \text{ curves.} \tag{3}$$

In lecture this week, we showed that if our model has has ϵ-covers of size $K_\epsilon \lesssim 1/\epsilon$, then the least squares estimator $\hat{\mu}$ converges at cube-root rate. That is, $\|\hat{\mu} - \mu^\star\|_{L_2(\mathrm{P_n})} \lesssim n^{-1/3}$ with high probability.

Exercise 1 *Explain why. This is just review.*

Labs earlier in the semester gave us empirical evidence that $\hat{\mu}$ converges at cube-root rate. Today, we'll prove it by finding an ϵ-cover of size $K_\epsilon \lesssim 1/\epsilon$.

As usual, we'll assume our covariates $X_1 \ldots X_n$ are in the unit interval $[0,1]$.

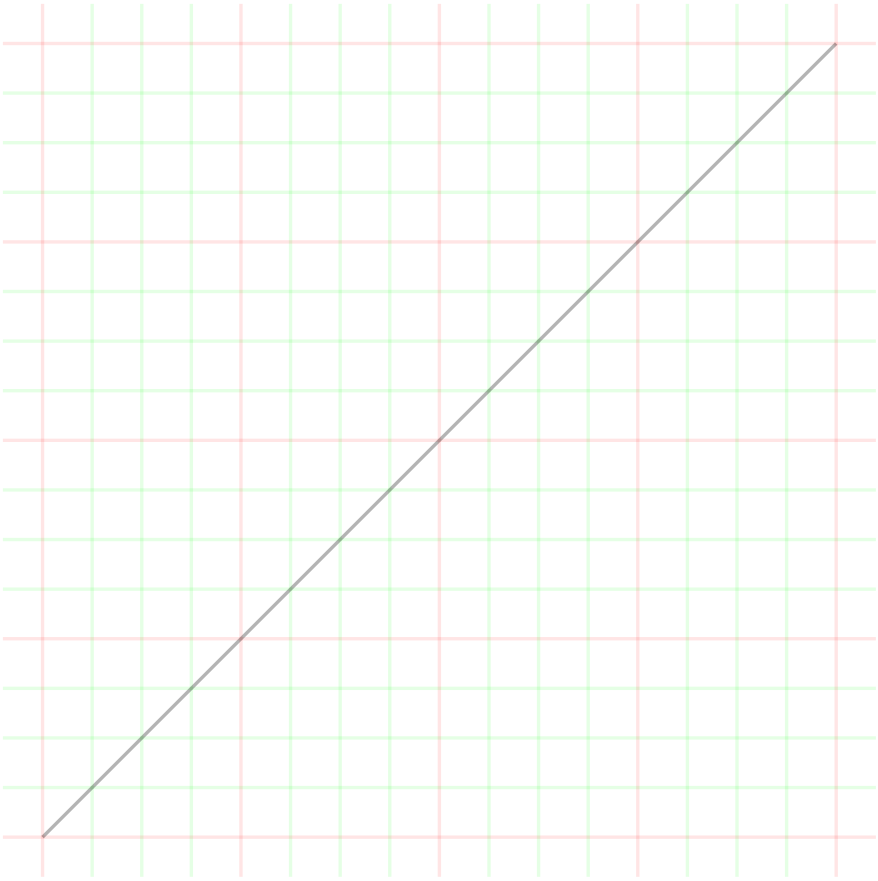
2 A Simplified Problem

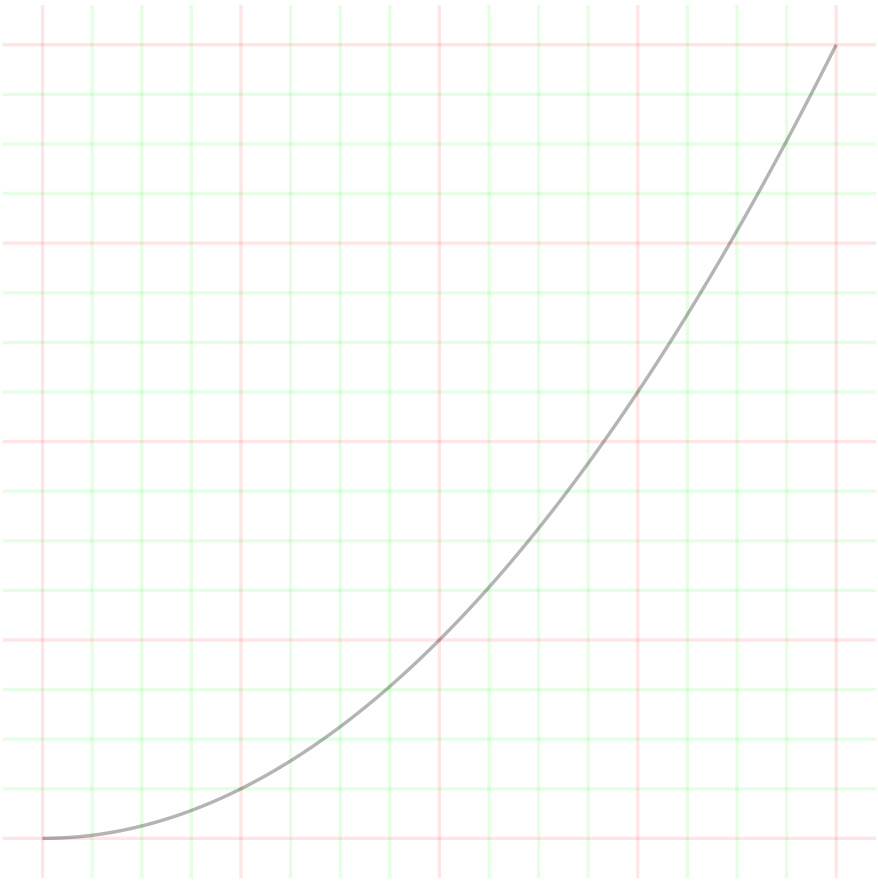
To keep things simple, we'll start by finding a cover for the subset of *increasing* curves in the bounded variation model (1). Because $\rho_{TV}(m) = m(1) - m(0)$ for increasing functions, this is just the set of increasing curves taking on values $m(x) \in [0, 1]$.

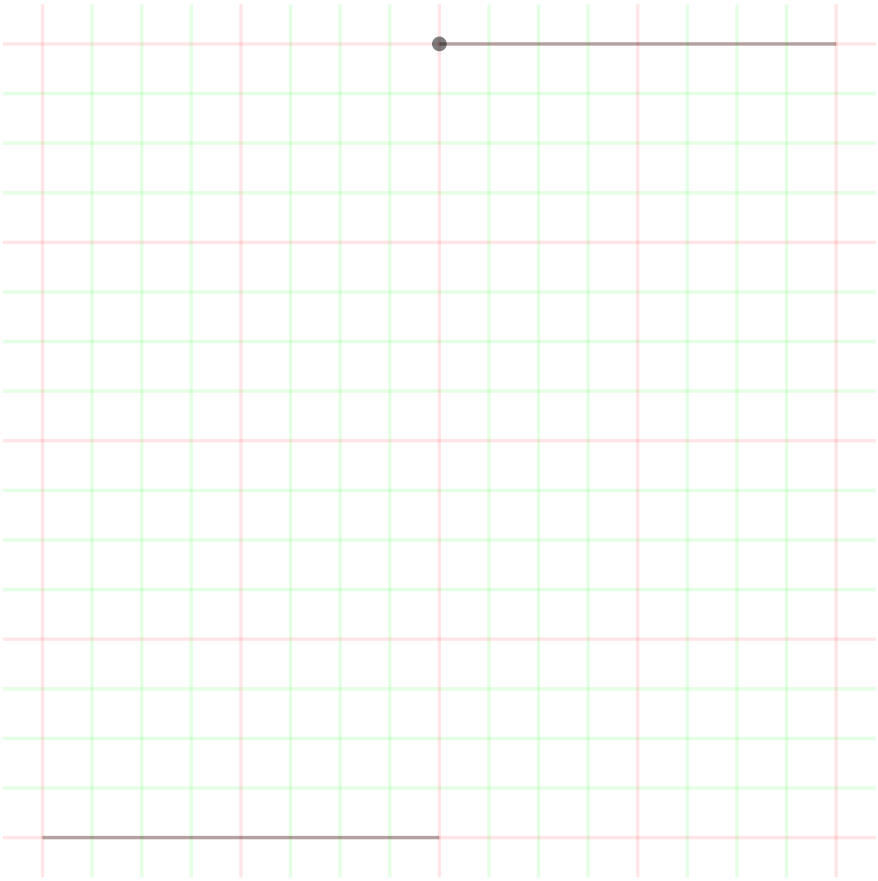
$$\mathcal{M} = \{\text{increasing } m \ : \ m(x) \in [0, 1]\}. \tag{4}$$

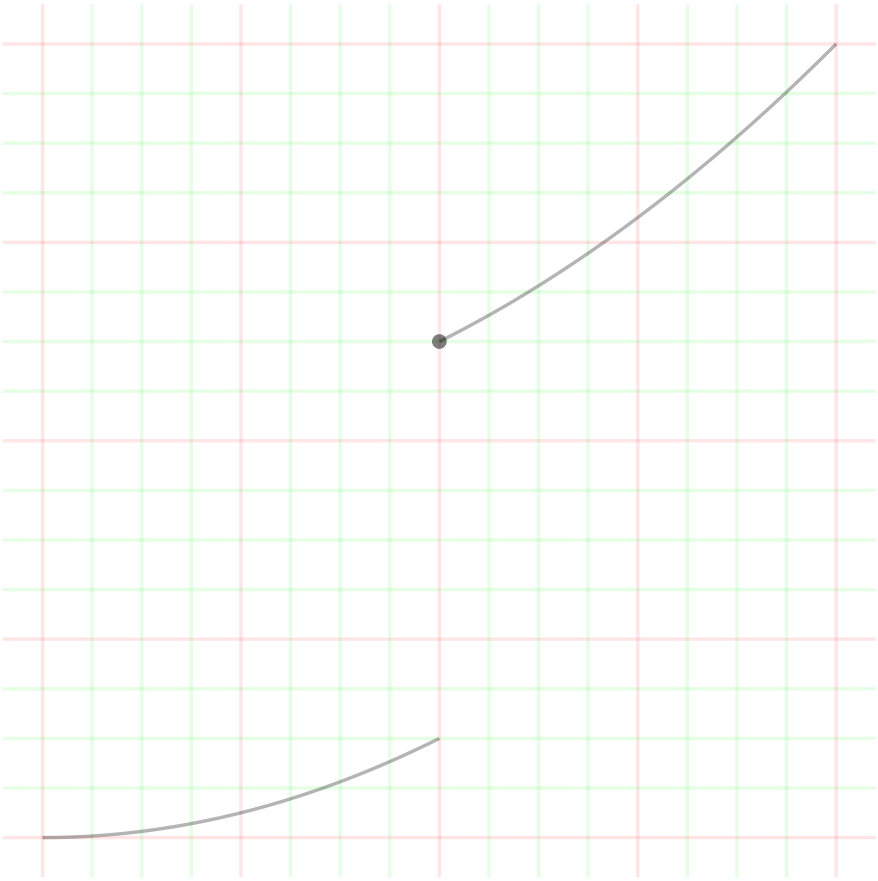
We'll use the same basic approach we used for Lipschitz functions in lecture: we'll *snap* our curve m to a piecewise-constant curve $\pi_\epsilon(m) = m_\epsilon$ where the value $m_\epsilon(x)$ at the grid points $x = 0, \epsilon, 2\epsilon, \ldots, (1/\epsilon) \cdot \epsilon = 1$ is $m(x)$ rounded to the nearest multiple of ϵ. Visually, that means that if we're drawing our curve on a grid with $\epsilon \times \epsilon$ squares, we snap (up or down) to the nearest horizontal grid line at each vertical grid line.

Exercise 2 *To get used to the idea, draw the curve $\pi_\epsilon(m)$ for the four curves m below. For each, do this for two values of ϵ: $\epsilon = 1/4$, corresponding to the red grid lines, and $\epsilon = 1/16$, corresponding to the green grid lines.*









Now let's get counting. This'll be our main exercise.

Exercise 3 *As a function of ϵ, bound the number of distinct curves $\pi_\epsilon(m)$ we get by snapping the set of all increasing curves with $m(x) \in [0,1]$. You may restrict your attention to the case that $\epsilon = 1/K$ for an integer K.*

Hint. Suppose we start with our pen hovering over the point $(x,y) = (0,0)$ and draw a piecewise constant increasing curve terminating at $(x,y) = (1, j\epsilon)$. The distance our pen moves when we draw it is $1 + j\epsilon$. And if we're drawing a 'snapped curve' $\pi_\epsilon(m)$, we move our pen only along the grid lines, taking $K = 1/\epsilon$ steps of length ϵ to the right and j steps of length ϵ upward. Each way of choosing j of these $K + j$ steps to move upward— or equivalently K of these $K + j$ steps to move rightward—yields a different curve $\pi_\epsilon(m)$. And there are $K + j$ **choose** $j = (K + j)!\ /\ (K!\ j!)$ such curves. The cover $\{\pi(m) : m \in \mathcal{M}\}$ contains each of these curves for every gridded endpoint $j\epsilon$ with $j \in 0 \ldots K$.

Hint. $(K+j)!\ /\ (K!\ j!) = (K+1)\dots(K+j)\ /\ j! \leq (2K)^j\ /\ j!$

Hint. Think about the Taylor series for e^x at $x = 2K$.

3 The Full Model

To count *all* the curves $\pi_\epsilon(m)$ for m in our bounded variation regression model $\mathcal{M}$ (1), we observe that we can write each such curve $m \in \mathcal{M}$ as the difference $m_+ - m_-$ of two increasing curves in $\mathcal{M}$. Where m increases, m_+ increases with it and $-m_-$ remains constant; where m decreases, $-m_-$ decreases with it and m_+ remains constant. This means we can come up with an ϵ-cover for all the curves in our model $\mathcal{M}$ using an ϵ-cover for the increasing ones.

Exercise 4 *Describe an ϵ-cover, or equivalently an ϵ-snapping map π_ϵ, for the bounded variation regression model $\mathcal{M}$ (1). State an upper bound on the number of curves it contains. Does your result imply an $n^{-1/3}$ rate of convergence? Why or why not?*

Hint. Don't get too fancy. If π is an ϵ-snapping map on the increasing functions in $\mathcal{M}$, is $\pi(m) = \pi(m_+) + \pi(m_-)$ an ϵ-snapping map on the whole model $\mathcal{M}$? How many distinct curves do we get if we snap $m \in \mathcal{M}$ like this?

4 Generalizing our Results

The monotone and bounded variation models we've been working with in this homework aren't quite the ones we've been using throughout the semester. We've imposed the additional constraint that $m(x) \in [0,1]$ for all x. We need these to get covering number bounds on the full model $\mathcal{M}$. After all, the set of all increasing curves and the set of curves with $\rho_{TV}(m) \leq 1$ both contain the constant curves $m(x) = \alpha$ for all $\alpha \in \mathbb{R}$, and we can't find an ϵ-cover of any finite size for those contant curves alone. However, *neighborhoods* in these models do not have this problem. Let's see what we can do.

We'll start by bounding the gaussian width of a neighborhood of zero, which is what we'd need to establish a rate of convergence in the admittedly implausible event that $\mu(x) = 0$.

Exercise 5 *Describe an ϵ-cover for a neighborhood of zero, $\mathcal{M}_s^0 = \{m \in \mathcal{M} : \|m\|_{L_2(\mathrm{P_n})} \leq s\}$ in (i) the set $\mathcal{M}$ of all increasing curves and (ii) the set $\mathcal{M}$ of all curves with $\rho_{TV}(m) \leq 1$. State an upper bound on the number of curves each contains. If it were true that $\mu(x) = 0$, would your result imply an $n^{-1/3}$ rate of convergence? Why or why not?*

Hint. Make very small changes to your answers to Exercises 3 and 4.

We've shown earlier in the semester that for models defined as balls in seminorms, like our bounded variation model $\mathcal{M} = \{m \ : \ \rho_{TV}(m) \leq 1\}$, the width of a neighborhood of an arbitrary point $\mu^\star \in \mathcal{M}$ can't be much bigger than the width of a neighborhood of zero. So Exercise 5 is really all you need to need to establish a rate of convergence irrespective of what μ is. But we have no such result for the monotone regression model. Let's approach this more directly.

Exercise 6 ***Optional.*** *Describe an ϵ-cover for a neighborhood of an arbitary curve $\mu^\star \in \mathcal{M}$, $\mathcal{M}_s = \{m \in \mathcal{M} \ : \ \|m - \mu^\star\|_{L_2(\mathrm{P_n})} \leq s\}$, in the set $\mathcal{M}$ of all increasing curves. Does your result imply an $n^{-1/3}$ rate of convergence? Why or why not?*