Convex Regression

January 30, 2025

In this homework, we'll work through the idea and implementation of *convex 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 Convexity

In some cases, we may believe that a curve we want to estimate is *convex*. A differentiable curve m is convex if its derivative is increasing.[1] More generally, a curve m is convex if all of its secants (line segments drawn from one point on the curve to another) lie above the curve, i.e., if for all points a and b it satisfies

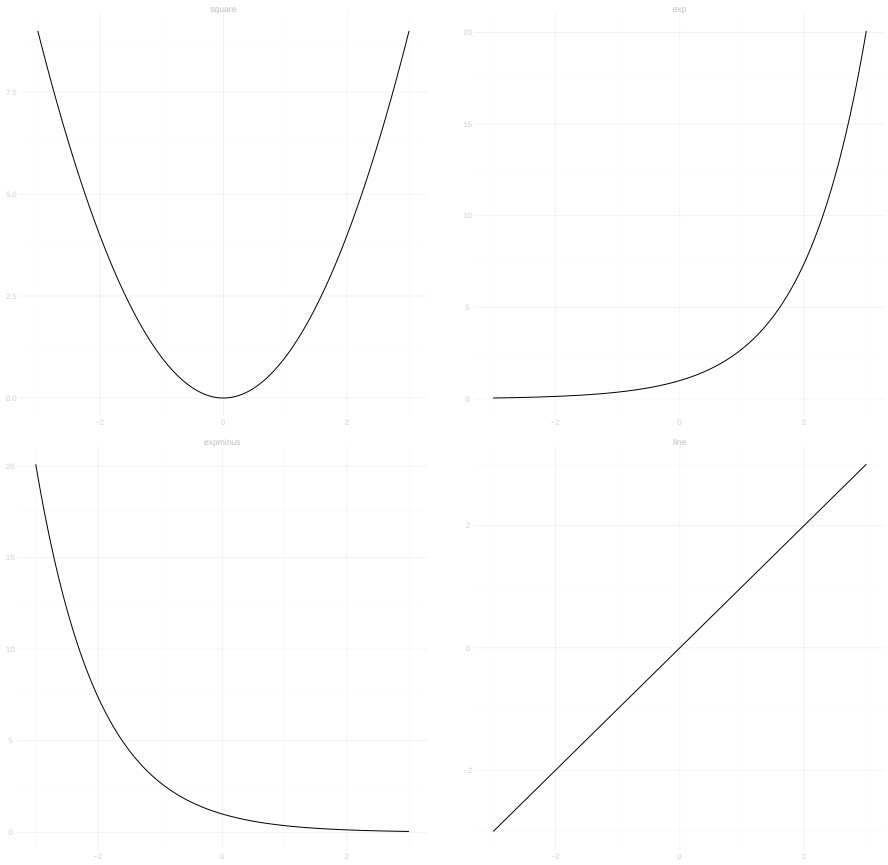
$$m\{(1-\lambda)a + \lambda b\} \leq (1-\lambda)m(a) + \lambda m(b) \quad \text{for all} \quad \lambda \in [0,1]. \tag{1}$$

This inequality is, in mathematical notation instead of visual language, exactly what we said about secants. The left side is the height of the curve at $x_\lambda = (1-\lambda)a + \lambda b$ and the right is the height of the secant connecting a to b at x_λ.

Characterizing points on secants. Keep in mind that any point x on the segment between a and b can be written in the form $x_\lambda = (1-\lambda)a + \lambda b$ for $\lambda \in [0,1]$. To do this, we simply solve the equation $x_\lambda = (1-\lambda)a + \lambda b$ for λ in terms of x, i.e., we take $\lambda = (x-a)/(b-a)$. This is pretty intuitive: λ is the fraction of the distance from a to b that we have to travel to get from a to x.

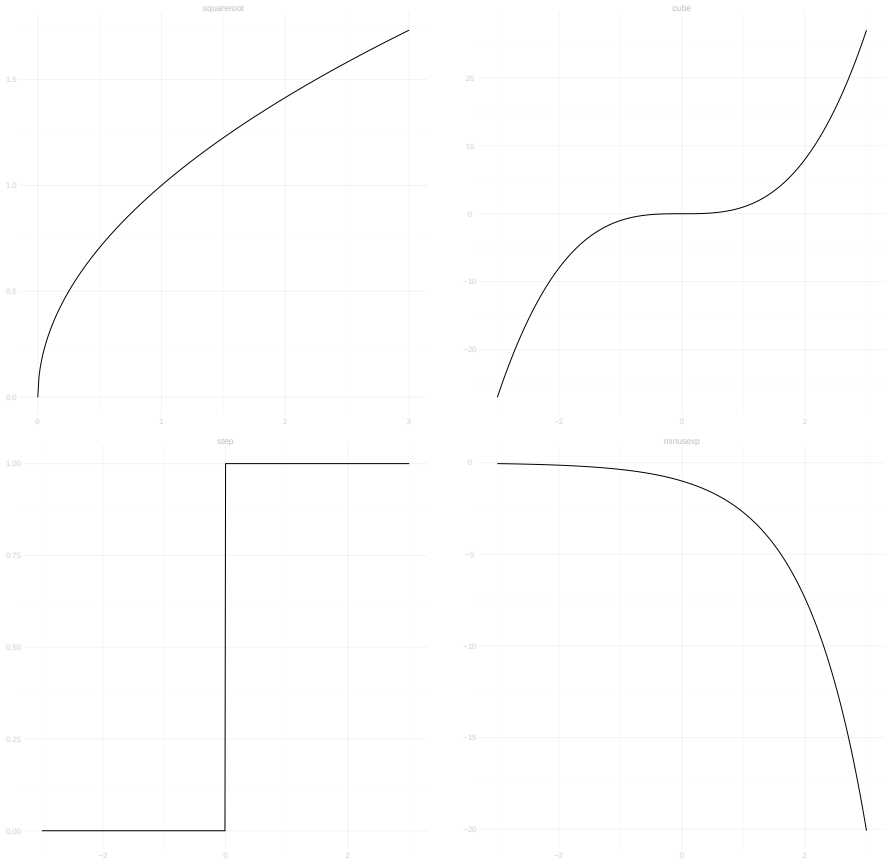
Examples. Here are some examples of convex curves.

1. $f(x) = x^2$
2. $f(x) = e^x$
3. $f(x) = e^{-x}$
4. $f(x) = x$


square
7.5
5.0
2.5
0.0
−2
0
2
exp
20
15
10
5
0
−2
0
2
expminus
20
15
10
5
0
−2
0
2
line
2
0
−2
−2
0
2

Here are a few curves that aren't convex.

1. $f(x) = \sqrt{x}$
2. $f(x) = (x - 1/2)^3$
3. $f(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$
4. $f(x) = -e^x$



Exercise 1 *On the eight plots above, draw a few secants. For the non-convex curves, make sure at least one is below the curve somewhere between the secant's endpoints.*

1.1 Differentiable Convex Functions

Now that we've got a sense of what's going on visually, let's argue that our more general definition based on (1) is consistent with the informal definition based

on derivatives I used in our first lecture.

Exercise 2 *Explain why, if a curve $m(x)$ is differentiable, it satisfies* (1) *if and only if its derivative $m'(x)$ is increasing.*

Hint. Here are two equivalent statements we can derive from (1) by taking $\lambda = (x-a)/(b-a)$ and $\lambda = 1-(x-b)/(b-a)$ respectively.

$$\begin{aligned} m(x) &\leq m(a) + \frac{m(b)-m(a)}{b-a}(x-a) \quad \text{for all } x \in [a,b] \\ m(x) &\leq m(b) + \frac{m(b)-m(a)}{b-a}(x-b) \quad \text{for all } x \in [a,b]. \end{aligned} \tag{2}$$

Rearranging, we get inequalities relating two slopes, one of which is the same in both cases.

$$\begin{aligned} \frac{m(x)-m(a)}{x-a} &\leq \frac{m(b)-m(a)}{b-a} && \text{for all } x \in [a,b] \\ \frac{m(b)-m(a)}{b-a} &\leq \frac{m(b)-m(x)}{b-x} = \frac{m(x)-m(b)}{x-b} && \text{for all } x \in [a,b]. \end{aligned} \tag{3}$$

What do these two equations together imply if we take $x \to a$ in the first and $x \to b$ in the second? This should help you show that convexity in the sense of (1) implies the increasingness of the derivative.

Another Hint. The mean value theorem tells us that, letting $x_\lambda = (1-\lambda)a + \lambda b$,

$$\begin{aligned} \frac{f(x_\lambda)-f(a)}{x_\lambda - a} &= f'(\tilde{a}) \quad \text{for some point} \quad \tilde{a} \in [a, x_\lambda] \\ \frac{f(b)-f(x_\lambda)}{b-x_\lambda} &= f'(\tilde{b}) \quad \text{for some point} \quad \tilde{b} \in [x_\lambda, b] \end{aligned} \tag{4}$$

If f' is increasing, how are these ratios related? And what, in terms of λ, a, and b, are their denominators? This should help you show that the increasingness of the derivative implies convexity in the sense of (1).

1.2 Convex Sets

There's a related notion of a *convex set*. We won't be using this for convex regression part of this homework, but it'll come up in lecture soon.

A convex set is a set that contains all line segments between points in it. That is, a set $\mathcal{S}$ is convex if and only if, for all points $a, b \in \mathcal{S}$, $(1-\lambda)a + \lambda b \in \mathcal{S}$ for all $\lambda \in [0,1]$. Here are a few examples.

In $1D$. A point, a line segment, or a line.

In $2D$. A filled-in triangle, square, or circle; the positive half-plane $\{(x,y) \in \mathbb{R}^2 : y \geq 0\}$; or the whole of $\mathbb{R}^2$.

Generally. A ball, the set $\{v : \rho(v) \leq r\}$, of any radius r in any seminorm ρ.

Here are a few sets that aren't convex.

In $1D$. Two points. Or the union of two disconnected intervals, e.g. $\{x : x \in [-1, 0] \text{ or } [1, 2]\}$.

In $2D$. A not-filled-in triangle, square, or circle.

In $3D$. A sphere, the set $\{v : \|v\| = r\}$, of any radius $r > 0$ in any norm.

Exercise 3 *Prove that a ball in a seminorm ρ is convex.*

Tip. Use the triangle inequality.

Exercise 4 *Using the norms we discussed in our Vector Spaces Homework, explain why that implies that the filled-in square $\{(x, y) : |x| \leq 1, |y| \leq 1\}$, circle $\{(x, y) : x^2 + y^2 \leq 1\}$, and diamond $\{(x, y) : |x| + |y| \leq 1\}$ are convex.*

Exercise 5 *Prove that a sphere of nonzero radius in any norm is not convex.*

Tip. Revisit the proof that seminorms are positive from the Vector Spaces Homework.

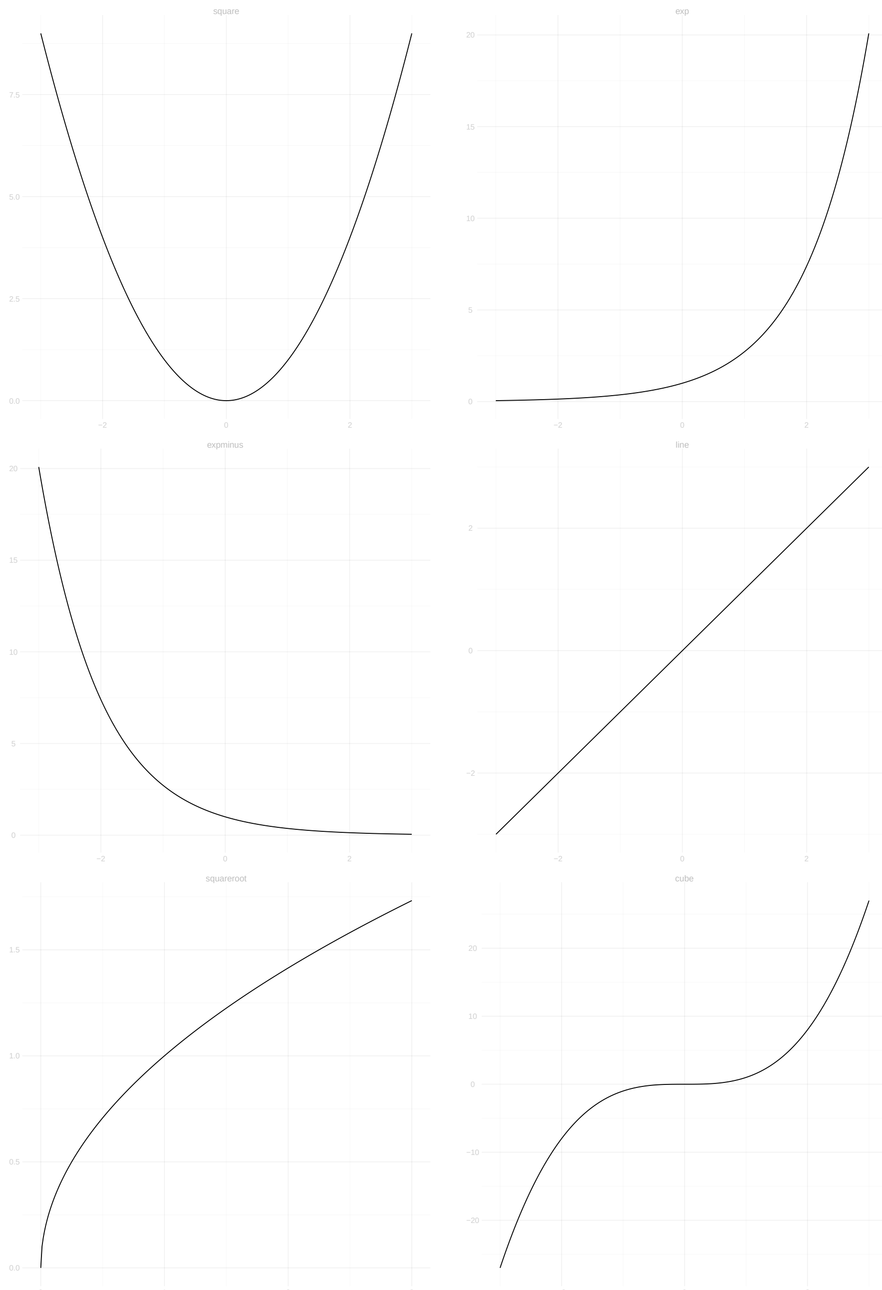
Exercise 6 *Draw, in 2D, a non-convex set that isn't included in the examples above.*

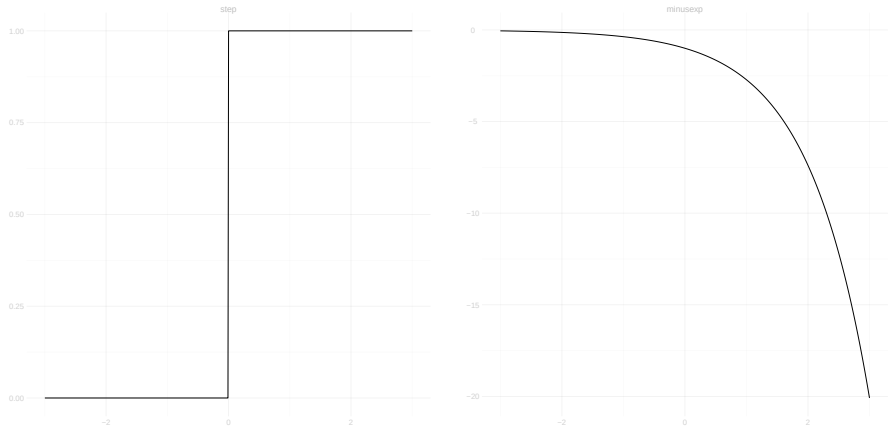
Exercise 7 *Explain why the intersection of two convex sets, i.e. the set of points that are in both of them, is a convex set.*

1.3 Convex functions have convex epigraphs

Here's another way of thinking about what convex functions look like. A function is convex if and only if its *epigraph*, the set of points on or above the curve, is convex. This is the definition of the epigraph of a function in mathematical notation. $\text{Epi}(f) = \{(x, y) : y \geq f(x)\}$.

Exercise 8 *On the plots below, fill in the epigraph.*


square
7.5
5.0
2.5
0.0
−2
0
2
exp
20
15
10
5
0
−2
0
2
expminus
20
15
10
5
0
−2
0
2
line
2
0
−2
−2
0
2
squareroot
1.5
1.0
0.5
0.0
0
1
2
3
cube
20
10
0
−10
−20
−2
0
2



Exercise 9 *(Optional) Explain why this epigraph-based definition is equivalent to the secant-based definition above in* (1)*. You don't have to give a formal proof.*

Tip. To show these definitions are equivalent, show that the convexity of a function's epigraph implies the convexity of the function and that the convexity of a function implies convexity of its epigraph. The latter part is a little harder. For intuition, try drawing a segment in a convex function's epigraph and the secant below it.

2 Convex Regression

Now that we've developed some intuition for what a convex function is, let's implement convex regression. That is, let's solve

$$\hat{\mu} = \underset{\text{convex } m:\mathbb{R}\to\mathbb{R}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\{Y_i - m(X_i)\}^2. \tag{5}$$

For this, we'll follow the same steps we used in the Monotone Regression Lab. We'll first solve a version of this problem for convex functions on the sample $\mathcal{X} = \{X_1 \dots X_n\}$, then extend our solution to the real line.

3 Fitting the Convex Regression Model

What does it mean for a function on $\mathcal{X}$ to be convex? We'll start with the same 'secant' definition we use for functions on $\mathbb{R}$, then forget about points that aren't in $\mathcal{X}$. That is, we'll say that $m : \mathcal{X} \to \mathbb{R}$ is convex if and only if

$$\lambda m(a) + (1-\lambda)m(b) \leq m\{(1-\lambda)a + \lambda b\}$$

whenever a, b, and $x_\lambda = (1-\lambda)a + \lambda b$ with $\lambda \in [0,1]$ are all in $\mathcal{X}$.

And we'll solve the restricted problem.

$$\hat{\mu}_{|\mathcal{X}} = \underset{\text{convex } m:\mathcal{X}\to\mathbb{R}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\{Y_i - m(X_i)\}^2. \tag{6}$$

Exercise 10 *Rewrite the optimization problem* (6) *more concretely in terms the values of* $X_1 \dots X_n$ *and* $m(X_1) \dots m(X_n)$.

Then translate it into a a constrained optimization over a vector $\vec{m}$*, so that, once you've solved for the optimal vector* $\vec{\mu}$*, you can express* $\hat{\mu}_{|\mathcal{X}}(X_1) \dots \hat{\mu}_{|\mathcal{X}}(X_n)$ *in terms of the elements of* $\vec{\mu}$ *(e.g. you might use* $\hat{\mu}_{|\mathcal{X}}(X_i) = \vec{\mu}_i$ *if* $X_1 \dots X_n$ *are distinct.) Try to do it so what you've written translates straightforwardly into CVXR code.*

Tips.

1. You should have a constraint for all triples (i, j, k) for which $X_i < X_j < X_k$. There is a smaller set of constraints that implies all of these, and we'll get there in Exercise 21, but it'll take some work. For now use the full set, like we did until the section 'Optional Exercise: Optimization' in the monotone regression lab.

2. 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.

Exercise 11 *Implement that optimization in* **R**. *That is, write an* **R** *function* **convexreg** *analogous to* **monotonereg** *from the monotone regression lab that solves* (6). *Then, from the eight distributions described below, sample* $n = 25$ *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* (6). *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 eight plots, labeling each with the signal used, as your solution to this exercise.*

We'll use as our signals μ the eight examples of convex and non-convex functions in Section 1. 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 $\sigma = 1/10$.

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) $a/b \le a'/b'$ and (ii) $ab' \le a'b$.

Exercise 12 *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 other model we've talked about could we do to fit the data better?*

Then, if there is a better model, use it and include the resulting plot.

3.1 Filling in the gaps

At this point, you have an estimator $\hat{\mu}_{|\mathcal{X}}$ that minimizes squared error among the convex functions $m : \mathcal{X} \to \mathbb{R}$. This lets us plot some isolated points. But we want a convex curve $\hat{\mu}(x)$ for $x \in [0, 1]$ and we want it to be the best-fitting such curve, i.e., we want the solution to (5).

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 $[X_n, 1]$.[2] This gives us a piecewise-linear solution to (6). First, we'll implement it. Then we'll verify that it is, in fact, a solution to (6).

Exercise 13 *Briefly explain why piecewise-constant extension would not give us a solution to* (6)*. A sentence or a sketch should do.*

Tip. Think about the examples from Section 1.

3.1.1 Implementation

Exercise 14 *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 Exercise 11.*

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

3.1.2 Verification

Exercise 15 *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 extension of the solution to* (6)*, $\hat{\mu} : \mathbb{R} \to \mathbb{R}$, is convex.*

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

3.2 Optimized Fitting

The downside of all this, from an implementation perspective, is that it involve *a lot* of constraints. The number of constraints is proportional to n^3. We can fix this. Ultimately, what we'll do is a lot like what we did to speed up monotone regression: we found a set of *local constraints* that determined whether a function was monotone. In particular, we found a way of establishing monotonicity by looking at pairs of neighboring observations instead of all pairs of observations.

3.3 Thinking locally about convexity

Let's think about whether we can use a *local properties* to determine whether a function is convex. By local property, I mean something you can check by looking only at small pieces of the function rather than the whole function all at once. For example, we know a function is increasing everywhere if it's increasing between n and $n+1$ for all integers n. This works more generally, if in place of the intervals $[n, n+1]$ we use any set of intervals that combine to cover the whole real line. And because a differentiable function is convex if and only if it has an increasing derivative, it follows that we can use this approach to determine whether a differentiable function is convex.

Let's try to generalize this. To start, it's worth observing that using exactly this approach won't work.

Exercise 16 *Describe a non-convex curve that is convex on the intervals* $[n, n+1]$ *for all integers* n*. Here, by convex on an interval, I mean that* (1) *holds for all points* a, b *in that interval.*

Tip. Look at the examples of non-convex curves above.

We can fix this by looking at *overlapping intervals* that cover the real line, for example, the intervals $[n-1, n+1]$. By overlapping, I mean that the endpoints of each interval are in the interior of (i.e. in but not endpoints of) some other interval. Our ultimate goal will be to show that a function is convex if it's convex on overlapping intervals that cover the real line. But to get the concepts down without messy arithmetic, let's start with something easier.

Exercise 17 *Show that if* $f(1) \leq \frac{1}{2}f(0) + \frac{1}{2}f(2)$ *and* $f(2) \leq \frac{1}{2}f(1) + \frac{1}{2}f(3)$*, then* $f(1) \leq \frac{2}{3}f(0) + \frac{1}{3}f(3)$*. Continue with this approach to show that* $f(1) \leq \frac{3}{4}f(0) + \frac{1}{4}f(4)$ *if, in addition,* $f(3) \leq \frac{1}{2}f(2) + \frac{1}{2}f(4)$*.*

It looks like there's a pattern here. If $f(n+1) \leq f(n) + f(n+2)$ for positive integers n, then $f(1) \leq \frac{n-1}{n}f(0) + \frac{1}{n}f(n)$. And because $1 = \frac{n-1}{n} \cdot 0 + \frac{1}{n} \cdot n$, this is an instance of our convexity-defining inequality (1) for $a = 0$ and $b = n$. If you're familiar with proof by induction, try the next exercise.

Exercise 18 *(Optional) Prove it! Use induction on* n*.*

The general case. If, for some increasing sequence $x_1 < x_2 < x_3 < \ldots < x_n$, a function f is convex on the overlapping intervals $[x_1, x_3]$, $[x_2, x_4]$, ..., $[x_{n-2}, x_n]$, then it's convex on the interval $[x_1, x_n]$. This is what we'll want when we're implementing our faster version of convex regression.

Exercise 19 *(Optional) Prove it!*

Tip. Start by showing that if f is convex on two intervals $[a, b]$ and $[b, c]$ and satisfies $f(b) \leq (1 - \lambda')f(a) + \lambda' f(c)$ for the value of $\lambda' \in [0, 1]$ for which $b = (1 - \lambda')a + \lambda' c$, then f is convex on $[a, c]$. To do this, it helps to observe that we can write $x \in [a, b]$ as $(1 - \lambda)a + \lambda b = (1 - \lambda)a + \lambda\{(1 - \lambda')a + \lambda' c\} = (1 - \lambda\lambda')a + \lambda\lambda' c$ and do something analogous for $x \in [b, c]$.

Tip. At some point in your argument, you'll probably want to take $a = x_1$, $b = x_3$, and $c = x_4$. To show that $f(x_3) \leq (1 - \lambda')f(x_1) + \lambda' f(x_4)$ for λ' such that $x_3 = (1 - \lambda')x_1 + \lambda' x_4$, you'll want to use the properties that $f(x_3) \leq (1 - \lambda'')f(x_2) + \lambda'' f(x_4)$ for λ'' such that $x_3 = (1 - \lambda'')x_2 + \lambda'' x_4$ and $f(x_2) \leq (1 - \lambda''')f(x_1) + \lambda''' f(x_3)$ for λ''' such that $x_2 = (1 - \lambda''')x_1 + \lambda''' x_3$.

Tip. The basic idea here is the same as the last exercise, but it's a bit messy. At least the way I did it. If you do want to try it, I recommend that you skip it on you first pass and come back to it when you've worked through the others.

3.4 Implementation

If our observations $X_1 \ldots X_n$ are distinct and sorted in increasing order, then $I_1 = [-\infty, X_2], I_2 = [X_1, X_3], I_3 = [X_2, X_4], \ldots, I_{n-1} = [X_{n-2}, X_n], I_n = [X_{n-1}, \infty]$ are overlapping intervals that cover the real line. A function $m : \mathbb{R} \to \mathbb{R}$ is convex if and only if it's convex on all of these intervals, i.e. if the restriction $m_{|I_j}$ is convex for all intervals I_j. And a function $m_{|\mathcal{X}} : \mathcal{X} \to \mathbb{R}$ is convex if and only the restriction to *the observations X_i in each interval*, i.e. the intersection $\mathcal{X}_j = \mathcal{X} \cap I_j$, is convex for all intervals I_j. What's cool about this is that there are either 2 or 3 observations in each of these sets ${\mathcal{X}_j}^3$, so — convexity on a set $\mathcal{X}_j$ being a property involving all *triples* in $\mathcal{X}_j$ — we get either 0 or 1 for each of these sets $\mathcal{X}_1 \ldots \mathcal{X}_n$. That means that, in total, we get no more than n constraints—in fact, we'll get $n - 2$.

Exercise 20 *Rewrite the optimization problem* (6) *more concretely in terms the values of $X_1 \ldots X_n$ and $m(X_1) \ldots m(X_n)$, this time using the $n - 2$ 'convexity on $\mathcal{X}_j$' constraints rather than the $\approx n^3$ constraints we used in Exercise 10.*

Then translate it into a a constrained optimization over a vector $\vec{m}$, so that, once you've solved for the optimal vector $\vec{\mu}$, you can express $\hat{\mu}_{|\mathcal{X}}(X_1) \ldots \hat{\mu}_{|\mathcal{X}}(X_n)$ in terms of the elements of $\vec{\mu}$ (e.g. you might use $\hat{\mu}_{|\mathcal{X}}(X_i) = \vec{\mu}_i$ if $X_1 \ldots X_n$ are distinct.) Try to do it so what you've written translates straightforwardly into CVXR code.

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.

Exercise 21 *Write a new version of* `convexreg` *that uses the constraints from Exercise 20. Then implement it and check that your solution agrees with the one you got using the all-triples constraints in Exercise 11. No need to turn in code, but you'll want this faster implementation for this next part.*

Repeat the fitting-and-plotting exercise from Exercise 11, but using sample size $n = 200$ *instead of* $n = 25$ *and plotting your solution's piecewise-linear extension* $\hat{\mu}$ *as a curve rather than the point predictions* $\hat{\mu}(X_1) \dots \hat{\mu}(X_n)$*. That is, for the eight distributions described below Exercise 11, sample* $n = 200$ *observations* $(X_1, Y_1) \dots (X_n, Y_n)$*, use your new version of* `convexreg` *together with your code from Exercise 14 to solve the convex regression problem* (5)*. Each time, plot the solution* $\hat{\mu}$ *(as a curve) on top of a scatter plot of the observations* (X_i, Y_i)*. Turn in those eight plots, labeling each with the signal used, as your solution to this exercise.*

4 Rates of Convergence

Now we've got three nonparametric regression models: monotone curves, bounded variation curves, and convex 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 22 *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, a bounded variation curve with budget $B = 1$*, and a convex curve. 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 23 *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.*