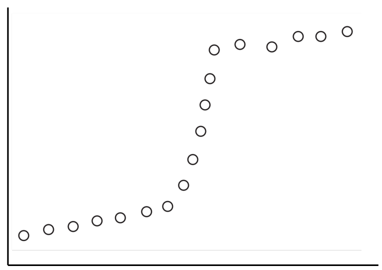
DS 320 Algorithms for Data Science **Lecture #14 Worksheet**
Spring 2023 Prof. Kira Goldner

Dynamic Programming II: Segmented Least Squares

The Problem

We are given a set of points $\{p_1 = (x_1, y_2), p_2 = (x_2, y_2), \ldots, p_n = (x_n, y_n)\}$ sorted by x-coordinate. Our goal is to fit a (segmented) line to P with least squares error.



What is "error" here? We use square error (SSE) from any line we use. That is, if our line is determined by slope a and y-intercept b, then our SSE would be

$$SSE = \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

Using calculus, we can derive that this is minimized when we set

$$a = \frac{n \sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)}{n \sum_i x_i^2 - (\sum_i x_i)^2} \qquad \text{and} \qquad b = \frac{\sum_i y_i - a \sum_i x_i}{n}.$$

But what if we can use as many segments as we want, just with a penalty c for each additional segment? How should we decide on the number of segments, and on what the segments should look like?

Our goal is to partition P into some C contiguous segments with minimal least squares error when there is a penalty c for each segment.

Making the Key Observation

The last point p_n belongs to a single segment which must begin somewhere. Where does it begin? In each case, what does the optimal solution look like?

Step 1: The Subproblem

Step 2: The Recurrence

Step 3: Prove that your recurrence is correct.

Step 4: State and prove your base cases.

Step 5: State how to solve the original problem.

Step 6: The Algorithm

Step 7: Running Time