Recall: Basis for cubic splines

Cubic splines basis: With 2 knots $\xi_1, \xi_2$:

- $h_1(X) = 1$
- $h_3(X) = X^2$
- $h_5(X) = (X - \xi_1)^3$
- $h_2(X) = X$
- $h_4(X) = X^3$
- $h_6(X) = (X - \xi_2)^3$

More generally, with $M$ knots, add $(X - \xi_1)^3, \ldots, (X - \xi_M)^3$.

Natural cubic splines basis: With $M$ knots

- $N_1(X) = 1$
- $N_2(X) = X$
- $N_{k+2}(X) = d_k(X) - d_{M-1}(x)$

where

$$d_k(X) = \frac{(X - \xi_k)^3 - (X - \xi_M)^3}{\xi_M - \xi_k}.$$

Example: Phoneme recognition

Example: Phoneme Recognition (ESL, Example 5.2.3)

$$X = X(f)$$
$$f = \text{frequency}.$$ 

$$\log \frac{P(\text{aa}|X)}{P(\text{ao}|X)} = \sum_{i=1}^{256} X(f_i) \beta_i = X^T \beta.$$ 

15 examples each of the phonemes “aa” and “ao” sampled from a total of 695 “aa” and 1022 “ao”.

Phoneme recognition (cont.)

Logistic regression coefficients, and smoothed version with natural cubic splines.

$$\beta(f) = \sum_{m=1}^{M} h_m(f) \theta_m = H \theta,$$

where $H$ is a $p \times M$ matrix of spline functions.

Now, note that

$$X^T \beta = X^T H \theta.$$

Letting $x^* = H^T x$, we can therefore fit the logistic regression on the smoothed inputs.
Work to do

- Write a function to construct natural cubic splines (can use a class if you want).
- Test your function:

![Graph](image.png)

- Construct the matrix $H \in \mathbb{R}^{p \times M}$ where $H_{ij} = h_j(f_i)$ as in the previous slide.
- Load the phoneme data. $X \in \mathbb{R}^{n \times p}, y \in \{0, 1\}^n$.
- Use a logistic regression on the transformed data $XH$ to predict the phonemes. Compute your prediction error.