COS 424
Homework #1
Due Tuesday, February 23rd

See the course website for important information about collaboration and late policies, as well as where and when to turn in assignments.

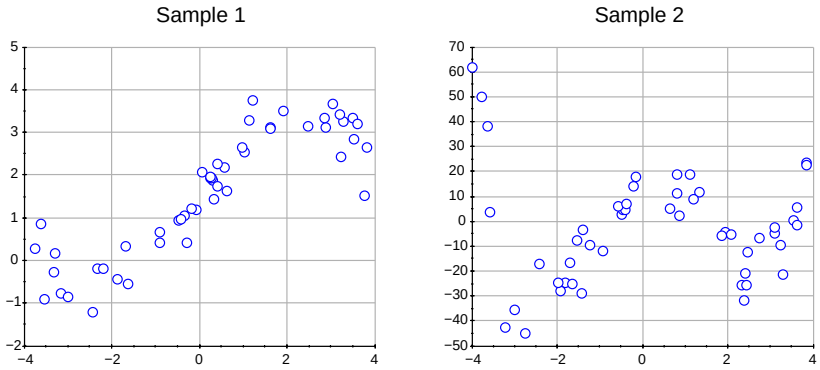
Data files

The questions make use of two data files available from
`http://www.cs.princeton.edu/courses/archive/spring10/cos424/w/hw1`.

```
hw1_sample2_train.txt
hw1_sample2_train.txt
```

Each data file contains $n = 50$ lines. Each line contains exactly 2 numbers representing the X and Y coordinates of points generated by adding noise to a secret function. The files `hw1_sample2_train.txt` and `hw1_sample2_train.txt` are only useful for the last question.

For verification, here is a graphical representation of both datasets.



Question 1 Implement linear least square curve fitting.

Provide plots showing the points and the fitted curve for both datasets using the following bases.

- Linear regression: $\Phi(x) = (1, x)$
- Cubic regression: $\Phi(x) = (1, x, x^2, x^3)$
- Cubic splines: $\Phi(x) = (1, x, x^2, x^3, [x+2]^3_+, [x]^3_+, [x-2]^3_+)$

The notation $[z]_+$ is a shorthand for $\max(0, z)$.

Question 2 Like polynomials, cubic splines often display strong variations outside of the domain of the input examples. Given a set of knots $r_1 < r_2 < \cdots < r_k$, we would like to construct a spline basis that always produces a piecewise function whose pieces are affine (polynomials of degree 1) when $x \leq r_1$ or $x \geq r_k$ and cubic (polynomials of degree 3) when $r_1 < x < r_k$.

Cubic splines can be written as

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \sum_{i=1}^{k} b_i [x - r_i]_+^3 \tag{1}$$

a) Write conditions on the parameters a_i and b_i expressing that the spline function has no cubic term when $x < r_1$ or $x > r_k$.

b) Derive a spline basis that embodies these conditions.

Hint: write $c_i = \sum_{j=1}^{i} b_j$ and rewrite expression (1) with the c_i instead of the b_i.

The resulting splines should have the form

$$f(x) = a_0 + a_1 x + a_2 x^2 + \sum_{i=1}^{k-1} c_i \phi_i(x) \tag{2}$$

c) Write further conditions on the coefficient a_i and c_i expressing that the spline function has no quadratic term when $x < r_1$ or $x > r_k$.

d) Derive a spline basis that embodies all these conditions, namely that the pieces of the resulting functions are affine when $x < r_1$ or $x > r_k$. Such splines are called *natural cubic splines*.

e) Count the number of dimensions and the number of constraints (continuity, continuous first derivatives, continuous second derivatives) to confirm that the basis dimension should indeed be k.

Question 3 We now consider natural splines with k evenly spaced knots such that $r_1 = -4$ and $r_k = 4$.

Use 5-fold cross-validation to determine the best k for each of the two samples.

Provide one plot for each dataset showing the average of the training and validation MSE obtained during k-fold cross-validation as a function of $k = 1 \ldots 8$.

Question 4 Download the two testing sets provided on the homework page and report the testing set MSE achieved by the models you have selected.