

## Homework 2

Due on November 22nd

We set up an experimental framework to study various aspects of overfitting. The input space is  $\mathcal{X} = [-1, 1]$  with uniform input probability density,  $P(x) = \frac{1}{2}$ . We consider the two models  $\mathcal{H}_2$  and  $\mathcal{H}_{10}$ . The target function is a polynomial of degree  $Q_f$ , which we write as  $f(x) = \sum_{q=0}^{Q_f} a_q L_q(x)$ , where  $L_q(x)$  are the Legendre polynomials. We use the Legendre polynomials because they are a convenient orthogonal basis for the polynomials on  $[-1, 1]$ .

The first two Legendre polynomials are  $L_0(x) = 1$ ,  $L_1(x) = x$ . The higher order Legendre polynomials are defined by the recursion:

$$L_k(x) = \frac{2k-1}{k} x L_{k-1}(x) - \frac{k-1}{k} L_{k-2}(x).$$

The data set is  $\mathcal{D} = (x_1, y_1), \dots, (x_N, y_N)$ , where  $y_n = f(x_n) + \sigma \epsilon_n$  and  $\epsilon_n$  are i.i.d. standard Normal random variables.

For a single experiment, with specified values for  $Q_f$ ,  $N$ ,  $\sigma$ , generate a random degree- $Q_f$  target function by selecting coefficients  $a_q$  independently from a standard Normal distribution, rescaling them so that  $\mathbb{E}_{\mathbf{a}, x}[f^2] = 1$ . Generate a data set, selecting  $x_1, \dots, x_N$  independently from  $P(x)$  and  $y_n = f(x_n) + \sigma \epsilon_n$ . Let  $g_2$  and  $g_{10}$  be the best fit hypotheses to the data from  $\mathcal{H}_2$  and  $\mathcal{H}_{10}$ , respectively, with respective out-of-sample errors  $E_{out}(g_2)$  and  $E_{out}(g_{10})$ .

- (a) Why do we normalize  $f$ ? [Hint: how would you interpret  $\sigma$ ?]
- (b) How can we obtain  $g_2$ ,  $g_{10}$ ? [Hint: pose the problem as linear regression]
- (c) How can we compute  $E_{out}$  analytically for a given  $g_{10}$ ?
- (d) Vary  $Q_f$ ,  $N$ ,  $\sigma$  and for each combination of parameters, run a large number of experiments, each time computing  $E_{out}(g_2)$  and  $E_{out}(g_{10})$ . Averaging these out-of-sample errors gives estimates of the expected out-of-sample error for the given learning scenario  $(Q_f, N, \sigma)$  using  $\mathcal{H}_2$  and  $\mathcal{H}_{10}$ . Let

$$\begin{aligned} E_{out}(\mathcal{H}_2) &= \text{average over experiments}(E_{out}(g_2)), \\ E_{out}(\mathcal{H}_{10}) &= \text{average over experiments}(E_{out}(g_{10})). \end{aligned}$$

Define the overfit measure  $E_{out}(\mathcal{H}_{10}) - E_{out}(\mathcal{H}_2)$ . When is the overfit measure significantly positive (i.e. overfitting is serious) as opposed to significantly negative? Try the choices  $Q_f \in \{1, 2, \dots, 100\}$ ,  $N \in \{20, 25, \dots, 120\}$ ,  $\sigma^2 \in \{0, 0.05, 0.1, \dots, 2\}$ . Explain your observations.

- (e) Why do we take the average over many experiments? Use the variance to select an acceptable number of experiments to average over.

Additional hints:

- The variance of the function  $f(x) = \sum_{q=0}^{Q_f} a_q L_q(x)$  on  $[-1, 1]$  is given by

$$\text{Var}(f(x)) = \sum_{q=0}^{Q_f} \frac{a_q^2}{2q+1}.$$

- The mean of the function  $f(x) = \sum_{q=0}^{Q_f} a_q L_q(x)$  on  $[-1, 1]$  is  $\mathbb{E}[f(x)] = a_0$  (all Legendre polynomials of degree  $> 0$  have mean 0 on  $[-1, 1]$ ).
- The variance of any random variable  $X$  is  $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ .