Homework 11
MATH 504

1.
- Reading from Elements of Statistical Learning (available in Course Documents)
  - Sec 5.1, 5.2 discuss spline regression and the general idea of using basis functions to calculate regressions. I followed roughly the notations and development of these sections in the lecture.
- Reading from Sauer
  - Sections 5.1.1, 5.1.2. Differentiation using finite difference approximations.
  - Sections 5.2.1, 5.2.2, 5.2.3. Integration using trapezoid rule and simpson's rule. Newton-Cotes formulas are generalizations of Riemann, trapezoid integration using higher order polynomials.
- Here is a nice webpage on spline regression in R, see section 4.5 therein.

  http://data.princeton.edu/R/linearModels.html

  The main point is that the R function **lm** is used in conjunction with an R function, e.g. **bs**, that selects a spline basis and then computes the associated model/design matrix. In otherwords, spline regression is viewed as a linear regression as we discussed in class and this view is made explicit by using **lm**.

2. The file **BoneMassData.txt** contains bone mass data for men and women at a variety of ages. This data comes from the book, "The Elements of Statistical Learning". In this problem, we will apply a spline regression to the dataset using predetermined knots. Specifically our regression model will be $y \sim S(x)$ were x is the age, y is the bone mass, and $S(x)$ is a cubic spline. For brevity consider only the samples taken from women and for simplicity consider two knots with $\zeta_1 = 15$ and $\zeta_2 = 20$ (the case of more knots is no different). Then our splines are determined by the parameters a_i, b_i, c_i, d_i for $i = 0, 1, 2$ where $S_i(x) = a_i + b_i x + c_i x^2 + d_i x^3$ and

$$S(x) = \begin{cases} S_0(x) & \text{if } x < \zeta_1 \\ S_1(x) & \text{if } x \in [\zeta_1, \zeta_2) \\ S_2(x) & \text{if } x \geq \zeta_2, \end{cases} \tag{1}$$

with the requirement that $S(x), S'(x), S''(x)$ be continuous at the two knots. Our goal is to find $S(x)$ that minimizes the sum of squared residuals

$$\sum_{i=1}^{N} (y_i - S(x_i))^2, \tag{2}$$

where (x_i, y_i) are the datapoints.

(a) Suppose that we can decompose any cubic spline $S(x)$ with knots at $\zeta = 15, 20$ as a linear combination of the functions $h_1(x), h_2(x), \ldots, h_D(x)$

for some D. That is, any cubic spline with knots at ζ_1, ζ_2 can be written as

$$S(x) = \sum_{j=1}^{D} \alpha_j h_j(x) \tag{3}$$

Show that the spline $S(x)$ that minimizes the sum of squared residuals is given by α defined by $\alpha = (B^T B)^{-1} B^T y$ where y is the vector of y_i values and B is a $N \times D$ matrix given by $B_{k\ell} = h_\ell(x_k)$. (We did this in class, I want you to go through the details.)

(b) Now show that D from (a) has value $D = 6$. (Again, we did this in class.)

(c) Show that the six functions $h_1(x) = 1$, $h_2(x) = x$, $h_3(x) = x^2$, $h_4(x) = x^3$, $h_5(x) = [x - \zeta_1]_+^3$, $h_6(x) = [x - \zeta_2]_+^3$ form a basis for all splines with knots at $\zeta = 15, 20$. To do this you must show (i) each $h_i(x)$ is a spline for the two knots, (ii) the $h_i(x)$ cannot be linearly combined to give the zero function, and (iii) each spline must be a linear combination of these functions. (Hint: show (i) and (ii) and then use the dimensionality of the spline space to show (iii))

(d) Now compute the regression spline $S(x)$ and plot it along with the data points to show the fit. (Typically it is better to do the actual fitting using a call to **lm** or using a spline regression package, but it's good to go through the process yourself at least once.)

3. Consider $f(x) = \mathrm{e}^x$. Note that $f'(0) = 1$. Consider the following two finite differences:

$$\frac{f(x+h) - f(x)}{h}. \tag{4}$$

$$\frac{f(x+h) - f(x-h)}{2h}. \tag{5}$$

For $h = 10^i$ with $i = -20, -19, -18, \ldots, -1, 0$ calculate both finite differences. For each h, determine how many digits in the finite difference estimate are correct (you know the true value of the derivative is 1). Note that .99991 is correct up to 4 digits since $.999999\cdots = 1$. Explain your results given finite differences and floating point error. DON'T FORGET TO SET **options(digits=16)**.

4. Let

$$F(x) = \int_0^x dz \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2}. \tag{6}$$

$F(x)$ is an important function - essentially the cdf of a normal random variable - that has no analytic formula and must be evaluated numerically. Write a function, **Fapprox**$(n, method)$ that approximates $F(\infty)$ by numerical integration. If **method** is set to the character string "reimann" or "trapezoid" use a Riemann sum and trapezoid rule method, respectively, with n grid points. If **method** is set to "useR", use the R integrate function with subdivisions set to n. (R's integrate function uses an adaptive grid method. In these methods, the grid is made dense in regions where the function varies, see Sauer for further details on such methods.) Since you cannot integrate to ∞, you must

pick some reasonable cutoff. We know that the true value is $F(\infty) = 1/2$. Consider approximating the integral using each of the three methods, try $n = 10, 100, 1000, 10000$ and compare accuracy.