Chapter 7 - Moving Beyond Linearity

1. $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-\xi)^3_+$
   $(x-\xi)^3_+ = \begin{cases} 0 & x \leq \xi; \\ (x-\xi)^3 & otherwise; \end{cases}$

   (a) $f_1(x) = a_1 + b_1 x + c_1 x^2 + d_1 x^3$
   For all $x \leq \xi$, $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$.
   Thus, $f_1(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

   (b) $f_2(x) = a_2 + b_2 x + c_2 x^2 + d_2 x^3$
   For all $x \leq \xi$, $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-\xi)^3$.
   $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x^3 - 3x^2\xi + 3x\xi^2 - \xi^3)$
   $f(x) = \beta_0 - \beta_4\xi^3 + (\beta_1 + 3\beta_4\xi^2)x + (\beta_2 - 3\beta_4\xi)x^2 + (\beta_3 + \beta_4)x^3$
   Thus, $f_2(x) = \beta_0 - \beta_4\xi^3 + (\beta_1 + 3\beta_4\xi^2)x + (\beta_2 - 3\beta_4\xi)x^2 + (\beta_3 + \beta_4)x^3$

   (c) $f_1(\xi) = \beta_0 + \beta_1\xi + \beta_2\xi^2 + \beta_3\xi^3$
   $f_2(\xi) = \beta_0 - \beta_4\xi^3 + (\beta_1 + 3\beta_4\xi^2)\xi + (\beta_2 - 3\beta_4\xi)\xi^2 + (\beta_3 + \beta_4)\xi^3$
   $f_2(\xi) = \beta_0 - \beta_4\xi^3 + \beta_1\xi + 3\beta_4\xi^3 + \beta_2\xi^2 - 3\beta_4\xi^3 + \beta_3\xi^3 + \beta_4\xi^3$
   $f_2(\xi) = \beta_0 + \beta_1\xi + \beta_2\xi^2 + \beta_3\xi^3$
   As $f_1(\xi) = f_2(\xi)$, $f(x)$ is continuous at ξ.

   (d) $f_1^{'}(x) = \beta_1 + 2\beta_2 x + 3\beta_3 x^2$
   $f_2^{'}(x) = \beta_1 + 3\beta_4\xi^2 + 2(\beta_2 - 3\beta_4\xi)x + 3(\beta_3 + \beta_4)x^2$
   $f_1^{'}(\xi) = \beta_1 + 2\beta_2\xi + 3\beta_3\xi^2$
   $f_2^{'}(\xi) = \beta_1 + 3\beta_4\xi^2 + 2(\beta_2 - 3\beta_4\xi)\xi + 3(\beta_3 + \beta_4)\xi^2$
   $f_2^{'}(\xi) = \beta_1 + 3\beta_4\xi^2 + 2\beta_2\xi - 6\beta_4\xi^2 + 3\beta_3\xi^2 + 3\beta_4\xi^2$
   $f_2^{'}(\xi) = \beta_1 + 2\beta_2\xi + 3\beta_3\xi^2$
   As $f_1^{'}(\xi) = f_2^{'}(\xi)$, $f^{'}(x)$ is continuous at ξ.

   (e) $f_1^{''}(x) = 2\beta_2 x + 6\beta_3 x$
   $f_2^{''}(x) = 2\beta_2 - 6\beta_4\xi + 6(\beta_3 + \beta_4)x$
   $f_1^{''}(\xi) = 2\beta_2\xi + 6\beta_3\xi$
   $f_2^{''}(\xi) = 2\beta_2 - 6\beta_4\xi + 6(\beta_3 + \beta_4)\xi$
   $f_2^{''}(\xi) = 2\beta_2 - 6\beta_4\xi + 6\beta_3\xi + 6\beta_4\xi$
   $f_2^{''}(\xi) = 2\beta_2\xi + 6\beta_3\xi$
   As $f_1^{''}(\xi) = f_2^{''}(\xi)$, $f^{''}(x)$ is continuous at ξ.

2. $\hat{g} = argmin(\sum\limits_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int [g^{(m)}(x)]^2 dx)$ where $g^{(m)}$ is the m^{th} derivative of g.

(a) As $\lambda \to \infty$, the smoothness penalty term is forced to be close to zero. Thus for $m = 0$, $g(x) \to 0$.

(b) As $\lambda \to \infty$, the smoothness penalty term is forced to be close to zero. Thus for $m = 1$, $g^1(x) \to 0$. This leads to $g(x)$ equaling a constant c.

(c) As $\lambda \to \infty$, the smoothness penalty term is forced to be close to zero. Thus for $m = 2$, $g^2(x) \to 0$. This leads to $g^1(x)$ equaling a constant c and $g(x)$ equaling a linear equation of the form $ax + b$.

(d) As $\lambda \to \infty$, the smoothness penalty term is forced to be close to zero. Thus for $m = 3$, $g^3(x) \to 0$. This leads to $g^2(x)$ equaling a constant c, $g^1(x)$ equaling a linear equation of the form $ax + b$, and $g(x)$ equaling a quadratic equation of the form $ax^2 + bx + c$.

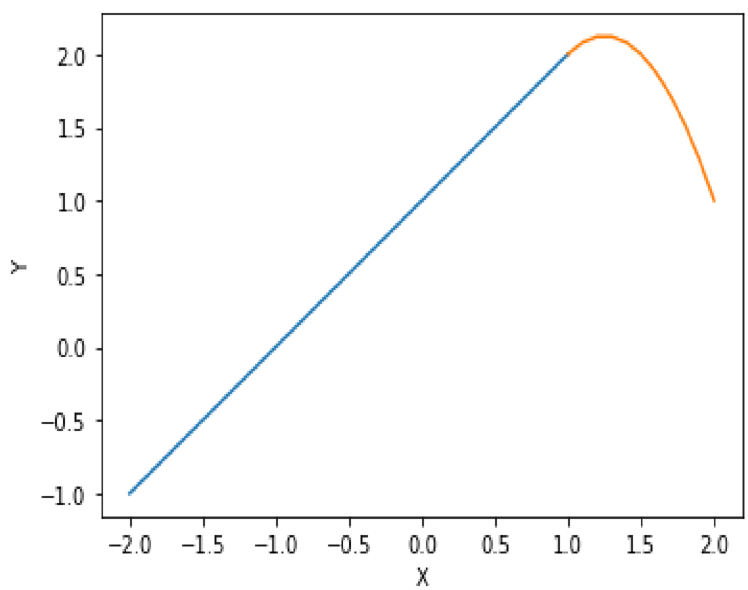
(e) As $\lambda = 0$, the smoothness penalty term is eliminated. Thus $g(x)$ will interpolate all data points and we will achieve a RSS of zero.

3. $b_1(X) = X$

$b_2(X) = (X-1)^2 I(X \geq 1)$

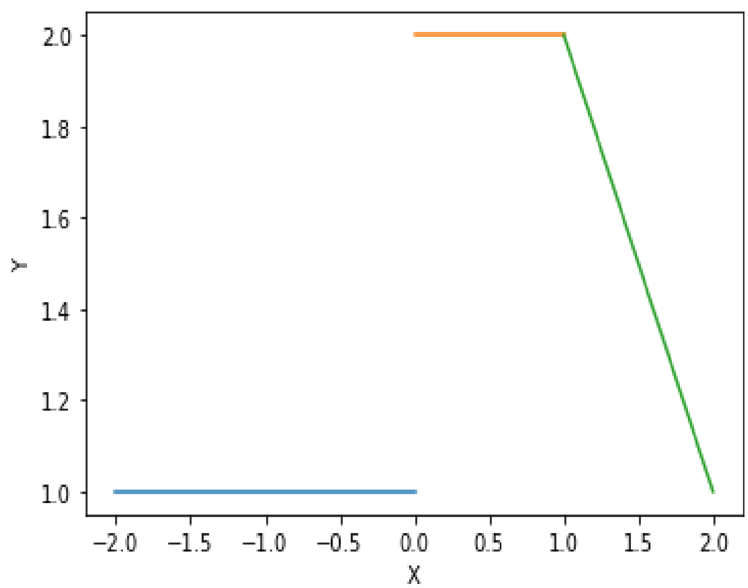
$Y = \beta_0 + \beta_1 b_1(X) + \beta_2 b_2(X) + \epsilon$

$$\hat{Y} = \begin{cases} 1 + X & X \leq 1; \\ 1 + X - 2(X-1)^2 & otherwise; \end{cases}$$



4. $b_1(X) = I(0 \leq X \leq 2) - (X-1)I(1 \leq X \leq 2)$
$b_2(X) = (X-3)I(3 \leq X \leq 4) + I(4 < X \leq 5)$
$Y = \beta_0 + \beta_1 b_1(X) + \beta_2 b_2(X) + \epsilon$

$$\hat{Y} = \begin{cases} 1 & X < 0, \\ 2 & 0 \leq X < 1, \\ 3 - X & 1 \leq X \leq 2, \\ 1 & 2 < X < 3, \\ 3X - 8 & 3 \leq X \leq 4, \\ 2 & 4 < X \leq 5, \\ 1 & X > 5, \end{cases}$$



5. (a) As $\lambda \to \infty$, $\hat{g}_2$ will have smaller training RSS than $\hat{g}_1$ as $\hat{g}_2$ is more flexible.

   (b) Test RSS depends on the true nature of the data. However, since $\hat{g}_2$ is more flexible, it has a greater chance of overfitting the data as compared to $\hat{g}_1$.

   (c) For $\lambda = 0$, $\hat{g}_1$ and $\hat{g}_2$ will be the same as the smoothness penalty term is eliminated.