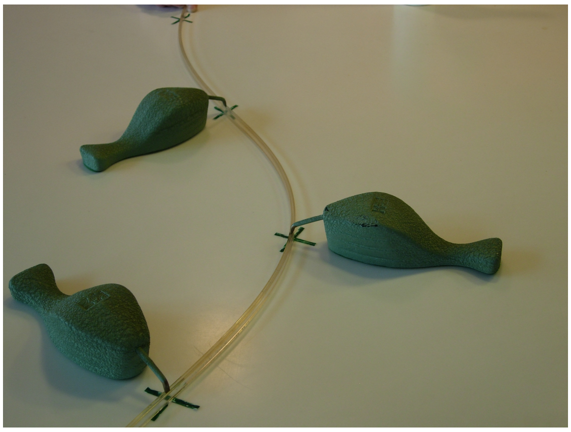
I.2.3. Cubic splines

In the last section we saw that Lagrange interpolation becomes impossible to use in practice if the number of points becomes large. Of course, the constraint we imposed, namely that the interpolating function be a polynomial of low degree, does not have any practical basis. It is simply mathematically convenient. Let's start again and consider how ship and airplane designers actually drew complicated curves before the days of computers. Here is a picture of a draughtsman's spline (taken from `http://pages.cs.wisc.edu/~deboor/draftspline.html` where you can also find a nice photo of such a spline in use)



It consists of a bendable but stiff strip held in position by a series of weights called ducks. We will try to make a mathematical model of such a device.

We begin again with points $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ in the plane. Again we are looking for a function $f(x)$ that goes through all these points. This time, we want to find the function that has the same shape as a real draughtsman's spline. We will imagine that the given points are the locations of the ducks.

Our first task is to identify a large class of functions that represent possible shapes for the spline. We will write down three conditions for a function $f(x)$ to be acceptable. Since the spline has no breaks in it the function $f(x)$ should be continuous. Moreover $f(x)$ should pass through the given points.

Condition 1: $f(x)$ is continuous and $f(x_i) = y_i$ for $i = 1, \ldots, n$.

The next condition reflects the assumption that the strip is stiff but bendable. If the strip were not stiff, say it were actually a rubber band that just is stretched between the ducks, then our resulting function would be a straight line between each duck location (x_i, y_i). At each duck location there would be a sharp bend in the function. In other words, even though the function itself would be continuous, the first derivative would be discontinuous at the duck locations. We will interpret the words "bendable but stiff" to mean that the first derivatives of $f(x)$ exist. This leads to our second condition.

Condition 2: The first derivative $f'(x)$ exists and is continuous everywhere, including each interior duck location x_i.

In between the duck locations we will assume that $f(x)$ is perfectly smooth and that higher derivatives behave nicely when we approach the duck locations from the right or the left. This leads to

Condition 3: For x in between the duck points x_i the higher order derivatives $f''(x), f'''(x), \ldots$ all exist and have left and right limits as x approaches each x_i.

In this condition we are allowing for the possibility that $f''(x)$ and higher order derivatives have a jump at the duck locations. This happens if the left and right limits are different.

The set of functions satisfying conditions 1, 2 and 3 are all the *possible* shapes of the spline. How do we decide which one of these shapes is the actual shape of the spline? To do this we need to invoke a bit of the physics of bendable strips. The bending energy $E[f]$ of a strip whose shape is described by the function f is given by the integral

$$E[f] = \int_{x_1}^{x_n} \left(f''(x)\right)^2 dx$$

The actual spline will relax into the shape that makes $E[f]$ as small as possible. Thus, among all the functions satisfying conditons 1, 2 and 3, we want to *choose the one that minimizes* $E[f]$.

This minimization problem is similiar to ones considered in calculus courses, except that instead of real numbers, the variables in this problem are *functions* f satisfying conditons 1, 2 and 3. In calculus, the minimum is calculated by "setting the derivative to zero." A similar procedure is described in the next section. Here is the result of that calculation: Let $F(x)$ be the function describing the shape that makes $E[f]$ as small as possible. In other words,

- $F(x)$ satisfies condtions 1, 2 and 3.
- If $f(x)$ also satisfies conditions 1, 2 and 3, then $E[F] \leq E[f]$.

Then, in addition to conditions 1, 2 and 3, $F(x)$ satisfies

Condition a: In each interval (x_i, x_{i+1}), the function $F(x)$ is a cubic polynomial. In other words, for each interval there are coefficients A_i, B_i, C_i and D_i such that $F(x) = A_i x^3 + B_i x^2 + C_i x + D_i$ for all x between x_i and x_{i+1}. The coefficients can be different for different intervals.

Condition b: The second derivative $F''(x)$ is continuous.

Condition c: When x is an endpoint (either x_1 or x_n) then $F''(x) = 0$

As we will see, there is exactly one function satisfying conditions 1, 2, 3, a, b and c.

I.2.4. The minimization procedure

In this section we explain the minimization procedure leading to a mathematical description of the shape of a spline. In other words, we show that if among all functions $f(x)$ satisfying conditions 1, 2 and 3, the function $F(x)$ is the one with $E[f]$ the smallest, then $F(x)$ also satisfies conditions a, b and c.

The idea is to assume that we have found $F(x)$ and then try to deduce what properties it must satisfy. There is actually a is a hidden assumption here — we are assuming that the minimizer $F(x)$ exists. This is not true for every minimization problem (think of minimizing the function $(x^2+1)^{-1}$ for $-\infty < x < \infty$). However the spline problem does have a minimizer, and we will leave out the step of proving it exists.

Given the minimizer $F(x)$ we want to wiggle it a little and consider functions of the form $F(x) + \epsilon h(x)$, where $h(x)$ is another function and ϵ be a number. We want to do this in such a way that for every ϵ, the function $F(x) + \epsilon h(x)$ still satisfies conditions 1, 2 and 3. Then we will be able to compare $E[F]$ with $E[F + \epsilon h]$. A little thought shows that functions of form $F(x) + \epsilon h(x)$ will satsify conditions 1, 2 and 3 for every value of ϵ if h satisfies

Condition 1': $h(x_i) = 0$ for $i = 1, \ldots, n$.

together with conditions 2 and 3 above.

Now, the minimization property of F says that each fixed function h satisfying 1', 2 and 3 the function of ϵ given by $E[F + \epsilon h]$ has a local minimum at $\epsilon = 0$. From Calculus we know that this implies that

$$\left.\frac{dE[F+\epsilon h]}{d\epsilon}\right|_{\epsilon=0} = 0. \tag{I.1}$$

Now we will actually compute this derivative with respect to ϵ and see what information we can get from the fact that it is zero for every choice of $h(x)$ satisfying conditions 1', 2 and 3. To simplify the presentation we will assume that there are only three points (x_1, y_1), (x_2, y_2) and (x_3, y_3). The goal of this computation is to establish that equation (I.1) can be rewritten as (I.2).

To begin, we compute

$$\begin{aligned} 0 = \left.\frac{dE[F+\epsilon h]}{d\epsilon}\right|_{\epsilon=0} &= \int_{x_1}^{x_3} \left.\frac{d(F''(x)+\epsilon h''(x))^2}{d\epsilon}\right|_{\epsilon=0} dx \\ &= \int_{x_1}^{x_3} 2\left.(F''(x)+\epsilon h''(x))h''(x)\right|_{\epsilon=0} dx \\ &= 2\int_{x_1}^{x_3} F''(x)h''(x)dx \\ &= 2\int_{x_1}^{x_2} F''(x)h''(x)dx + 2\int_{x_2}^{x_3} F''(x)h''(x)dx \end{aligned}$$

We divide by 2 and integrate by parts in each integral. This gives

$$0 = F''(x)h'(x)\Big|_{x=x_1}^{x=x_2} - \int_{x_1}^{x_2} F'''(x)h'(x)dx + F''(x)h'(x)\Big|_{x=x_2}^{x=x_3} - \int_{x_2}^{x_3} F'''(x)h'(x)dx$$

In each boundary term we have to take into account the possibility that $F''(x)$ is not continuous across the points x_i. Thus we have to use the appropriate limit from the left or the right. So, for the first boundary term

$$F''(x)h'(x)\Big|_{x=x_1}^{x=x_2} = F''(x_2-)h'(x_2) - F''(x_1+)h'(x_1)$$

Notice that since $h'(x)$ is continuous across each x_i we need not distinguish the limits from the left and the right. Expanding and combining the boundary terms we get

$$\begin{aligned} 0 = &-F''(x_1+)h'(x_1) + \big(F''(x_2-) - F''(x_2+)\big)h'(x_2) + F''(x_3-)h'(x_3) \\ &- \int_{x_1}^{x_2} F'''(x)h'(x)dx - \int_{x_2}^{x_3} F'''(x)h'(x)dx \end{aligned}$$

Now we integrate by parts again. This time the boundary terms all vanish because $h(x_i) = 0$ for every i. Thus we end up with the equation

$$\begin{aligned} 0 = &-F''(x_1+)h'(x_1) + \big(F''(x_2-) - F''(x_2+)\big)h'(x_2) + F''(x_3-)h'(x_3) \\ &+ \int_{x_1}^{x_2} F''''(x)h(x)dx - \int_{x_2}^{x_3} F''''(x)h(x)dx \end{aligned} \tag{I.2}$$

as desired.

Recall that this equation has to be true for every choice of h satisfying conditions 1', 2 and 3. For different choices of $h(x)$ we can extract different pieces of information about the minimizer $F(x)$.

To start, we can choose h that is zero everywhere except in the open interval (x_1, x_2). For all such h we then obtain $0 = \int_{x_1}^{x_2} F''''(x)h(x)dx$. This can only happen if

$$F''''(x) = 0 \quad \text{for} \quad x_1 < x < x_2$$

Thus we conclude that the fourth derivative $F''''(x)$ is zero in the interval (x_1, x_2).

Once we know that $F''''(x) = 0$ in the interval (x_1, x_2), then by integrating both sides we can conclude that $F'''(x)$ is constant. Integrating again, we find $F''(x)$ is a linear polynomial. By integrating four times, we see that $F(x)$ is a cubic polynomial in that interval. When doing the integrals, we must not extend the domain of integration over the boundary point x_2 since $F''''(x)$ may not exist (let alone by zero) there.

Similarly $F''''(x)$ must also vanish in the interval (x_2, x_3), so $F(x)$ is a (possibly different) cubic polynomial in the interval (x_2, x_3).

(An aside: to understand better why the polynomials might be different in the intervals (x_1, x_2) and (x_3, x_4) consider the function $g(x)$ (unrelated to the spline problem) given by

$$g(x) = \begin{cases} 0 & \text{for } x_1 < x < x_2 \\ 1 & \text{for } x_2 < x < x_3 \end{cases}$$

Then $g'(x) = 0$ in each interval, and an integration tells us that g is constant in each interval. However, $g'(x_2)$ does not exist, and the constants are different.)

We have established that $F(x)$ satisfies condition a.

Now that we know that $F''''(x)$ vanishes in each interval, we can return to (I.2) and write it as

$$0 = -F''(x_1+)h'(x_1) + \big(F''(x_2-) - F''(x_2+)\big)h'(x_2) + F''(x_3-)h'(x_3)$$

Now choose $h(x)$ with $h'(x_1) = 1$ and $h'(x_2) = h'(x_3) = 0$. Then the equation reads

$$F''(x_1+) = 0$$

Similarly, choosing $h(x)$ with $h'(x_3) = 1$ and $h'(x_1) = h'(x_2) = 0$ we obtain

$$F''(x_3-) = 0$$

This establishes condition c.

Finally choosing $h(x)$ with $h'(x_2) = 1$ and $h'(x_1) = h'(x_3) = 0$ we obtain

$$F''(x_2-) - F''(x_2+) = 0$$

In other words, F'' must be continuous across the interior duck position. Thus shows that condition b holds, and the derivation is complete.

This calculation is easily generalized to the case where there are n duck positions $x_1, \ldots, x_n$.

A reference for this material is *Essentials of numerical analysis, with pocket calculator demonstrations,* by Henrici.

I.2.5. The linear equations for cubic splines

The following section contains an alternative more efficient version.

Let us now turn this description into a system of linear equations. In each interval (x_i, x_{i+1}), for $i = 1, \ldots n-1$, $f(x)$ is given by a cubic polynomial $p_i(x)$ which we can write in the form

$$p_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i$$