CPSC 303: ENERGY IN CUBIC SPLINES, POWER SERIES AS ALGORITHMS, AND THE INITIAL VALUE PROBLEM

JOEL FRIEDMAN

Contents

Disclaimer: The material may sketchy and/or contain errors, which I will elaborate upon and/or correct in class. For those not in CPSC 303: use this material at your own risk...

The field of splines is quite remarkable and deep; furthermore the various notions of "boundary conditions" (including "initial conditions") has connections to ODE's and three-term recurrences (where we usually specify "initial conditions"). The course textbook [A&G] by Ascher and Greif covers splines in Chapter 11, although the focus is on algorithms. The goal of this note is to make some additional, clarifying remarks on splines.

The main new ideas we introduce beyond [A&G] are (1) the "energy minimizing" nature of cubic splines, (2) the algorithmic implications of this "energy," (3) energy dissipation and the local nature of splines (this is mentioned briefly [A&G]), and

Date: Tuesday 2nd April, 2024, at 12:50(get rid of time in final version).
Research supported in part by an NSERC grant.

(4) the "initial value" boundary conditions. In the process thereof we get to review power series of matrices, diagonal dominance, and similarity—which is presumably covered in CPSC 302—but is easily within our grasp since we have already discussed the p-norms of matrices.

1. Splines are "Energy" Minimizers

One way to describe splines is to fix real numbers

$$A = x_0 < x_1 < \ldots < x_n = B,$$

and to fix any $y_0, \ldots, y_n \in \mathbb{R}$, and then to consider all functions

$$\mathcal{U} = \mathcal{U}_{\mathbf{t},\mathbf{y}} \stackrel{\text{def}}{=} \Big\{ u \in C^2[A,B] \;\Big|\; u(x_i) = y_i \text{ for all } i \Big\}, \tag{1}$$

where $C^2[A,B]$ denotes the set of functions $[A,B] \to \mathbb{R}$ that are twice continuously differentiable on $[A,B]$ (for the endpoints A, B we take the derivative only from one side). One can prove that there is a unique function, $v = v(x) \in \mathcal{U}$ that minimizes the "energy"

$$\mathcal{E}(u) = \mathcal{E}_{\text{SecondDeriv}}(u) \stackrel{\text{def}}{=} \int_A^B \big(u''(x)\big)^2 \, dx \tag{2}$$

over all functions in $\mathcal{U} = \mathcal{U}_{\mathbf{t},\mathbf{y}}$. It is the unique function $v \in \mathcal{U}$ such that

(1) v restricted to any interval $[x_i, x_{i+1}]$ is a cubic polynomial, i.e., of degree at most 3, and
(2) $v''(A) = v''(B) = 0$ (the derivative $u''(A)$ exists "from the right," and $v''(B)$ from the left).

It is called the *cubic spline* (through (x_i, y_i) for $i = 0, \ldots, n$) with *free boundary conditions*, meaning $v''(x_0) = v''(x_n) = 0$; any such spline with free boundary conditions is called a *natural spline*.

The *free* in "free boundary conditions" means that no "boundary conditions" are specified for $\mathcal{U}$, and that the resulting minimizer, v, has $v''(x_0) = v''(x_n) = 0$. The standard example of a "boundary condition," where the "boundary" refers to the endpoints $A = x_0$ and $B = x_n$, are the *clamped* boundary conditions where one considers for some $\alpha, \beta \in \mathbb{R}$:

$$\mathcal{U}_{\alpha,\beta}^{\text{clamped}} \stackrel{\text{def}}{=} \{u \in \mathcal{U} \mid u'(x_0) = \alpha, u'(x_n) = \beta\},$$

and it turns out that there is a unique energy minimizer in this subset of $\mathcal{U}$.

Some of you have seen the clamped conditions in 2-dimensional drawing programs, where certain arrows (curves, etc.) routines require you to specify the two endpoints of the arrow, i.e., $u(A)$ and $u(B)$, plus the angle that the curve makes with the endpoints, which is information regarding $u'(A)$ and $u'(B)$. The difference is that here u has two-dimensional values; also, often there are default values for $u'(A)$ and $u'(B)$, and mostly likely you don't have the freedom to specify arbitrary values for both $u'(A)$ and $u'(B)$.

Not surprisingly, if you consider clamp only the "left" side of the spline, i.e., you consider the subset of $\mathcal{U}$ given by

$$\mathcal{U}_{\alpha}^{\text{left clamped}} \stackrel{\text{def}}{=} \{u \in \mathcal{U} \mid u'(x_0) = \alpha\}$$

for some $\alpha \in \mathbb{R}$, then one can prove that there is a unique minimizer, v, and at the "right" endpoint $x_n = B$ we have $v''(x_n) = 0$ (the free condition on "the right").

We have used the term "energy" with no particular motivation from physics. If you don't like thinking of $\mathcal{E}$ above as "energy," you can also think of it as "a desirable property to be minimized," similar to the "least squares" measure and related measures of "goodness of fit" in data modeling.

Because we are considering the "energy" (2) as integrating over $(u''(t))^2$, it should not come as a surprise that our algorithm to determine the minimizer v (under any of the above boundary conditions) focuses on v''. In particular, if set $c_i = v''(x_i)/2$ for $i = 0, \dots, n$, then it turns out that v'' is the unique piecewise-linear function (i.e., the "bending line" or, in [A&G], "broken line") with $v''(x_i) = 2c_i$; our algorithms first solve for $c_0, \dots, c_n$ (as a tri-diagonal system), and from there the rest of v is easily determined. [The reason for the factor of 2 in $v''(x_i) = 2c_i$, as well as other details, will be explained in the next section.]

The idea of "energy minimization" is studied in more general terms in physics (e.g., Lagrangian mechanics) and in math (e.g., the calculus of variations). We will comment about some of the subtle points of this area after discussing some computational aspects in [A&G], Section 11.3.

Typically one thinks of splines as modeling some curve $f\colon [A, B] \to \mathbb{R}$, such as the profile of an airplane wing or a car (or, more often, a surface in $\mathbb{R}^3$). One assumes that one can evaluate f at any value; so we are relatively free to choose n and a sequence of distinct (usually increasing) values $\mathbf{t} = (x_0, \dots, x_n)$, and we modify $\mathcal{U}_{\mathbf{t},\mathbf{y}}$ in (1) by substituting $f(x_i)$ for y_i everywhere, i.e., we set

$$\mathcal{U} = \mathcal{U}_{f;\mathbf{t}} \stackrel{\text{def}}{=} \Big\{ u \in C^2[A, B] \;\Big|\; u(x_i) = f(x_i) \text{ for all } i \Big\}. \tag{3}$$

At times we assume that some values of f' can be determined with reasonable accuracy, and at times we do not: mostly we will only consider f' at the endpoints; for example, to model f with the clamped boundary we would consider the subset of $\mathcal{U} = \mathcal{U}_{f;\mathbf{t}}$ given by

$$\Big\{ u \in \mathcal{U}_{f;\mathbf{t}} \;\Big|\; u'(x_0) = f'(x_0) \text{ and } u'(x_n) = f'(x_n) \Big\}.$$

[When we can reliably measure f' anywhere, and when f'''' is reasonably bounded, then Section 11.2 of [A&G] explains the advantages of piecewise Hermite interpolation.]

2. Notation in Section 11.3 of [A&G]

Discussing splines can get awkward without some carefully chosen notation. Here we review the notation in Section 11.3 of [A&G].

Recall the definition of $C^2[A, B]$ from the beginning of Section 1.

Definition 2.1. Let

$$A = x_0 < x_1 < \cdots < x_n = B$$

be real numbers. A *cubic spline* with *endpoints* $A = x_0$ and $B = x_n$, and *break points* $x_1, \dots, x_{n-1}$ is any function, $v \in C^2[A, B]$ whose restriction to each interval $[x_i, x_{i+1}]$, for $i = 0, \dots, n-1$, is a cubic polynomial (i.e., of degree at most 3).

With notation as in the definition above, [A&G] introduces the following notation: we set

$$h_0 = x_1 - x_0, \ \dots \ , h_{n-1} = x_n - x_{n-1}.$$

[We will use similar notation h_i when we study differential equations.] We use the notation

$$s_i(x) = a_i + b_i(x-x_i) + c_i(x-x_i)^2 + d_i(x-x_i)^3 \quad \text{for } x_i \le x \le x_{i+1} \tag{4}$$

for the cubic pieces of $v(s)$. This notation is used in Section 11.3 of [A&G], although it is already introduced—with minor differences—in Section 11.2. Throughout Section 11.3 we are fixing a function $f\colon [A,B] \to \mathbb{R}$ whose values we can examine, and we restrict our splines to the set $\mathcal{U} = \mathcal{U}_{f;\mathbf{t}}$ in (3).

One reason that this notation is clever is that differentiating (4) we easily verify that

$$\begin{aligned} a_i =& v(x_i) = s_i(x_i) = s_{i-1}(x_i), \\ b_i =& v'(x_i) = s_i'(x_i) = s_{i-1}'(x_i), \\ 2c_i =& v''(x_i) = s_i''(x_i) = s_{i-1}''(x_i), \quad \text{and} \\ 6d_i =& s_i'''(x_i) = s_{i-1}'''(x_i). \end{aligned}$$

for $i = 0, \dots, n-1$, where for $i = 0$ we ignore $s_{i-1}, s_{i-1}', s_{i-1}''$ (or imagine that the spline is extended to a $x_{-1} < x_0 = A$); the equalities on the $6d_i$ line hold because s_i''' is a constant; the equalities on the $a_i, b_i, 2c_i$ lines hold because of the existence of v, v', v'' at the break points. In the above notation we have

$$s_i''(x) = 2c_i + 6d_i(x - x_i)$$

and it follows that v'' is a bending-line, and hence v''' doesn't exist at the break-points, in the sense that its left- and right-limits exist but are not generally the same (i.e., we cannot expect that $d_i = d_{i+1}$).

As mentioned in Section 1, it becomes crucial to introduce c_n via

$$c_n = v''(x_n)/2 = s_{n-1}''(x_n)/2;$$

it is also helpful to introduce a_n, b_n, d_n via

$$a_n = v(x_n) = s_{n-1}(x_n), \ b_n = v'(x_n) = s_{n-1}'(x_n),$$

$$d_n = v''(x_n)/6 = s_{n-1}'''(x_n)/6 = s_{n-1}'''(x_{n-1})/6.$$

The energy $v''(t)$ is therefore the unique bending line with values

$$v''(x_i) = 2c_i;$$

it follows that once we determine $c_0, \dots, c_n$, we can solve for d_i with $i \le n-1$ (which is all we need to determine the $s_i(x)$ and therefore to determine v) via

$$d_i = \frac{c_{i+1} - c_i}{x_{i+1} - x_i} = \frac{c_{i+1} - c_i}{h_i}.$$

Of course, the values of the a_i are known immediately from f, i.e.,

$$a_i = f(x_i).$$

Of course, the above notation chooses to expand each s_i by its left endpoint, x_i, rather than its right endpoint, x_{i+1}; one could equivalently use right endpoints everywhere.

Notice that the textbook uses $[a,b]$ instead of $[A,B]$, and this may cause some confusion between a, b and the a_i, b_i above.

3. Boundary Conditions

We easily see that conditions on the a_i, b_i, c_i, d_i give $4n-2$ linear equations: indeed, we have $v(x_i) = f(x_i)$ for $i = 0, \ldots, n$, giving $n+1$ equations, and the continuity conditions for v, v', v'' at the $n-1$ break points give total of

$$n + 1 + 3(n-1) = 4n - 2.$$

Since there are $4n$ variables in total, **one can hope** that one can specify two additional constraints, i.e., linear equations, to get a unique spline with given values $v(x_i) = f(x_i)$ for $i = 0, \ldots, n$. The two additional constraints that are typically used are "boundary conditions" in the sense that they are conditions at $x = A = x_0$ and $x = B = x_n$, or nearby.

There are four main interesting boundary conditions:

Initial conditions: we specify $v'(x_0)$ and $v''(x_0)$. The textbook does not discuss this, but this is the analog of initial conditions in three-term recurrences and initial conditions in ODE's. It turns out to be easy to see that there exists a unique spline with these boundary conditions; we will discuss this in the next section.

Free boundary conditions: we impose $v''(x_0) = v''(x_n)$; this spline turns out to be the minimizer of the energy

$$\mathcal{E}(u) \stackrel{\text{def}}{=} \int_A^B |u''(x)|^2 \, dx$$

over all splines. This gives the *natural spline*.

Clamped conditions: we specify $v'(x_0)$ and $v'(x_n)$; if we are modeling a function f where we can compute f', then we should take $v'(x_0) = f'(x_0)$ and $v'(x_n) = f'(x_n)$. This spline turns out to be the minimizer of the energy

$$\mathcal{E}(v) \stackrel{\text{def}}{=} \int_A^B |v''(x)|^2 \, dx$$

over all splines such that $v'(x_0)$ and $v'(x_n)$ have the values we have specified for them.

Not-a-knot: Which specifies v''' should be continuous across x_1 and x_{n-1}. This turns out to give a unique spline and does not require any derivative information.

In class we consider the case $n = 1$; the solution is a unique linear function, i.e., the solution to the differential equation $v''(x) = 0$, except that the not-a-knot is hard to interpret (without an "expert on the empty set").

4. Spline Algorithms

We now describe the idea behind many spline algorithms. We have already determined $a_i = f(x_i)$ for $i = 0, \ldots, n-1$, so it suffices to solve for the b_i, c_i, d_i.

4.1. Initial Value Problems.

In the initial value problem we choose $\beta, \gamma \in \mathbb{R}$ and specify $v'(x_0) = \beta$ and $v''(x_0) = \gamma$; we claim that there is a unique spline satisfying

these two conditions. To see this, we first solve for a_0, b_0, c_0, d_0:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 1 & h_0 & h_0^2 & h_0^3 \end{bmatrix} \begin{bmatrix} a_0 \\ b_0 \\ c_0 \\ d_0 \end{bmatrix} = \begin{bmatrix} f(x_0) \\ \beta \\ \gamma \\ f(x_1) \end{bmatrix} \tag{5}$$

Once we have determined a_0, b_0, c_0, d_0, and therefore $s_0(x)$, we determine a_1, b_1, c_1, d_1:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 1 & h_1 & h_1^2 & h_1^3 \end{bmatrix} \begin{bmatrix} a_1 \\ b_1 \\ c_1 \\ d_1 \end{bmatrix} = \begin{bmatrix} f(x_1) \\ s_0'(x_1) \\ s_0''(x_1) \\ f(x_2) \end{bmatrix}$$

where we use the fact that v', v'' are required to be continuious across the break point x_1, and hence

$$s_0'(x_1) = s_1'(x_1), \quad s_0''(x_1) = s_1''(x_1).$$

We similarly solve for a_i, b_i, c_i, d_i for $i = 2, 3, \ldots, n-1$.

Notice that if we know $f'(x_0)$ and $f''(x_0)$ we can use these values as β and γ respectively in (**??**); in practice this isn't usually a good way to model f.

The real reason why initial conditions are interesting is that:

(1) they are analogous to "initial conditions" in recurrences (which we have covered) and ODE's (which we will cover), and
(2) these illustrate "shooting methods" (see the homework and class discussion) which are useful in ODE's: here we choose two (or more) pairs of values for β, γ, see what the two splines looks like at $x_n = B$, and then choose a third (or fourth, fifth, etc.) pair to get a desired condition at $x_n = B$.

Unfortunately, there are so many amazing properties of splines that we tend to have little use for "shooting methods" and "initial value conditions" in practice.

4.2. **Power Series and Diagonal Domiance.** At this point we break to discuss power series in linear algebra. These would be covered in CPSC 302 and linear algebra courses beyond the minimum CPSC 303 prerequisite of one term of linear algebra. Fortunately, in CPSC 303 we have covered matrix norms.

There is a fundamental difference between the series

$$1 + (1/2)^2 + (1/3)^2 + \cdots \quad \text{which happens to equal } \pi^2/6$$

and

$$1 - (1/2) + (1/2)^2 - (1/2)^3 + \cdots \quad \text{which happens to equal } 2/3$$

The second series converges quickly, since when we truncate it after the first m terms, the remainder is

$$(1/2)^m + (1/2)^{m+1} + \cdots = 2^{-m+1}$$

which converges to 0 exponentially as a function of m. The remainder of the first series is

$$(m+1)^{-2} + (m+2)^{-2} + \cdots = 1/m + O(1/m^2),$$

where $O(1/n^2)$ means a term of "order" $1/m^2$; this means that the first series does give an "algorithm" to approximately compute $\pi^2/6$, but to get 20 digits of accuracy you would have to add roughly 10^{20} terms of this series. By contrast,

the first series converges exponentially quickly to 2, and could be used as a toy "algorithm" to compute 2/3.

[It is true that

$$1-(1/2)+(1/2)^2-(1/2)^3+\cdots = \big(1-(1/2)\big)+\big((1/2)^2-(1/2)^3\big)+\cdots = 1/2+1/8+1/32+\cdots$$

which does give the base 2 expression for 2/3.]

Many similar remarks are true for the power series of matrices that arise in this section. Let us summarize what we need to know.

To discuss the matrix power series, first note that

$$(1+x)^{-1} = 1 - x + x^2 - x^3 + \cdots$$

is a convergent power series for $|x| < 1$; what this means is that if we set

$$(1+x)^{-1} = 1 - x + x^2 - x^3 + \cdots + (-x)^{m-1} + t_m,$$

or equivalently define

$$t_m = (1+x)^{-1} - \big(1 - x + x^2 - x^3 + \cdots + (-x)^{m-1}\big),$$

then we have

$$|t_m| \leq \frac{|x|}{1-|x|}.$$

There are a number of ways of proving this bound on t_m, but if you trust that the above power series converges, and trust "absolute value' distributes over (countably infinite) sums, then you have values

$$\begin{aligned} |t_m| &= \big|(-x)^m + (-x)^{m+1} + (-x)^{m+2} + \cdots\big| \\ &\leq |x|^m + |x|^{m+1} + \cdots \\ &\leq |x| \,/\, (1-|x|). \end{aligned}$$

Now let us explain the analog for matrices.

So fix an $n \times n$ matrix A and a p such that $\|A\|_p < 1$ (in this section we will only be interested in the case $p = \infty$). We claim that $I - A$ exists, and is given by the convergent *matrix* power series

$$(I+A)^{-1} = I - A + A^2 - A^3 + \cdots$$

where the right-hand-side is a *convergent matrix power series*: everything really works the way it does for absolute value and real numbers, you just have to get used to the p-norm (really only $p = \infty$ in this section) and its analogous properties:

$$\|\alpha A\|_p = |\alpha| \, \|A\|^p, \quad \|AB\|_p \leq \|A\|_p \, \|B\|_p, \quad \|A^m\|_p \leq \|A\|_p^m, \quad \|A+B\|_p \leq \|A\|_p + \|B\|_p$$

(for all square matrices A, B, real α, positive integer m, $p \geq 1$ or $p = \infty$) this properties are not difficult to prove (see the exercises), and have only slight differences (e.g., $|x^m| = |x|^m$ for $x \in \mathbb{R}$ and positive integers m, as opposed to the inequality $\|A^m\|_p \leq \|A\|_p^m$). So if we truncate this series after m terms, writing

$$(I+A)^{-1} = I - A + A^2 - A^3 + \cdots + (-A)^{m-1} + T_m,$$

then

$$(-A)^m + (-A)^{m+1} + (-A)^{m+2} + \cdots,$$

and let us now see that this term is bounded by

$$\|T_m\|_p = \Big\|(-A)^m + (-A)^{m+1} + (-A)^{m+2} + \cdots \Big\|_p$$
$$\le \|A\|_p^m + \|A\|_p^{m+1} + \cdots = \frac{\|A\|_p}{1-\|A\|_p}.$$

We are about to encounter (in (6)) square matrices N with $\|N\|_\infty \le 1$ and an expression

$$\big(I + N/2\big)^{-1},$$

for which we have a similar expansion

$$\big(I + N/2\big)^{-1} = I + (-N/2) + (-N/2)^2 + \cdots + (-N/2)^{m-1} + U_m$$

where we therefore have (writing $A = N/2$ in the above discussion)

$$\|U_m\|_\infty \le 2^{-m+1},$$

which decays exponentially as a function of m. This shows that $(I + N/2)^{-1}$ is well approximated by the first n terms, and hence (1) one can *understand the "effect"* $(I + N/2)^{-1}$ from its "first few terms," and (2) one even has an algorithm for approximating $(I + N/2)^{-1}$ when N is a *sparse* matrix (see [A&G] and/or CPSC 302).

4.3. **Neighbours and Power Series.** The reason we used the (mildly awkward choice of) letter N is to suggest the fact that the matrix N is only non-zero between neighbours: the matrix N associated to (6) has the property that its (i,j) entry is only nonzero for $|i-j| = 1$, i.e., for immediate neighbours. It follows that N^2 is only nonzero for $|i-j| \le 2$, and more precisely only for $|i-j| = 0, 2$. Similarly N^3 is only nonzero for $|i-j| \le 3$, and more precisely only for $|i-j| = 1, 3$. It follows that the approximation

$$\big(I + N/2\big)^{-1} \approx I + (-N/2) + (-N/2)^2 + \cdots + (-N/2)^{m-1}$$

reflects only an "$m-1$ nearest neighbour effect," in that the matrix on the right-hand-side is only nonzero in (i,j)-entries with $|i-j| \le m-1$.

4.4. **The Free Boundary Problem: Executive Summary.** We now point out that a lot of boundary conditions yield equations in $c_0, \ldots, c_{n-1}, c_n$, and that once we know the c_i's, we can easily determine the $s_i(x)$—we already have seen that we can determine the $d_i = (c_{i+1} - c_i)/h_i$ and that $a_i = f(x_i)$ that we can solve for the b_i and d_i once we know (for all i) the value of c_i.

Since the derivations are all give in [A&G], Section 11.3, we only summarize the equations and examine their intuitive meaning.

4.4.1. *The Equations for the c_i.* The key formula in [A&G] that describes our algorithm to find cubic splines involves the equation

$$\frac{h_{i-1}}{h_{i-1}+h_i} c_{i-1} + 2c_i + \frac{h_i}{h_{i-1}+h_i} c_{i+1} = 3f[x_{i-1}, x_i, x_{i+1}], \quad i = 1, \ldots, n-1 \tag{6}$$

which is essentially equation (11.5) there; this gives us a set of $n-1$ equations the $n+1$ variables $c_0, \ldots, c_{n-1}, c_n$. To solve the free boundary conditions, we simply set $c_0 = c_n = 0$ and solve the resulting $(n-1) \times (n-1)$ system of equations; the reason this is not hard to solve is presumably discussed in CPSC 302 (see Section 5.6 of [A&G]) or any math course that discusses the LU-decomposition;

the *diagonally dominance* of the system implies that the LU-decomposition works (without permutations), and that the system has a unique solution.

4.4.2. *Localization or Dissipation of the Energy of f.* The above equation shows that we set A to be the off diagonal entries of the above tridiagonal system, then the resulting matrix is of the form $2I + N$, where N has row sums at most 1. It follows that $\|N\|_\infty \leq 1$ (with equality iff $n \geq 3$, since for $n = 1, 2$, for any i, (6) at least one of c_{i-1}, c_i, c_{i+1} is known and zero). It follows that $\|N/2\|_\infty \leq 1$, hence one has a convergent power series

$$2\big(I + (N/2)\big)^{-1} = 2\big(I - (N/2) + (N/2)^2 - (N/2)^3 + \cdots\big);$$

when we truncate this power series after the $(N/2)^m$ term, we get an error term whose ∞-norm is

$$\leq 2 \sum_{i \geq m+1} \big\|(N/2)^i\big\|_\infty \leq 2(2^{-m-1} + 2^{-m-2} + \cdots) = 2^{-m+1},$$

which therefore decays exponentially as a function of m. Since the (i, j)-th component of N is zero unless $|i - j| \leq 1$, we easily see that the (i, j)-th component of A^m is zero unless $|i - j| \leq m$. This implies a sort of "localization" of the linear system; in other words, for fixed i, the "energy of f at i," i.e., the second order difference $3f[x_{i-1}, x_i, x_{i+1}]$, has an exponentially decaying effect on values of c_j as a function of $|i - j|$.

4.4.3. *The Equations for b_i.* We then solve for the b_i via

$$b_i = f[x_i, x_{i+1}] - h_i(2c_i + c_{i+1})/3$$

(see equations (11.4a) and (11.4b)).

4.5. **Clamped Boundary Conditions.** For clamped boundary conditions, the c_0 and c_n are unknowns. However [A&G], middle of page 343 (Section 11.3), explain that clamping on the left yields the equation

$$2c_0 + c_1 = 3\frac{f[x_0, x_1] - f'(x_0)}{h_0},$$

and since

$$\frac{f[x_0, x_1] - f'(x_0)}{h_0} = \frac{f[x_0, x_1] - f[x_0, x_0]}{h_0} = f[x_0, x_0, x_1]$$

this can be written as

$$2c_0 + c_1 = 3f[x_0, x_0, x_1];$$

similarly the clamping on the right yields the analogous

$$c_{n-1} + 2c_n = 3\frac{f'(x_n) - f[x_{n-1}, x_n]}{h_n} = 3f[x_{n-1}, x_{n-1}, x_n].$$

It follows that with clamping on the left or right or on both sides, each clamping introduces one addition constraint, with the same 2 on the diagonal entry, and the same off diagonal entry sum of at most 1. It follows that the exact same remarks regarding the now $n \times n$ system (if we clamp on one side) or $(n+1) \times (n+1)$ system (if we clamp on both sides) holds as they did in the free boundary case.

5. Why Splines Miminize Energy

In this section we prove a theorem that indicates that splines minimize the energy, as promised in Section 1. However, this theorem assumes that (1) an energy minimizer exists, and (2) that this minimizer is sufficiently differentiable; we will discuss these assumptions in the next section.

Similar to the definition of $C^2[A,B]$ in Section 1, for real $x_0 < x_1$, we use $C^4[x_0,x_1]$ to denote the set of functions $[x_0,x_1] \to \mathbb{R}$ that are four times continuously differentiable on $[x_0,x_1]$ (for the endpoints x_0, x_1 we take the derivative only from one side).

Theorem 5.1. *Let $x_0 < x_1$ be real and let*

$$\mathcal{U} \subset C^2[x_0,x_1]$$

be any subset such that for any $u \in \mathcal{U}$ we have $u+g \in \mathcal{U}$ for any $g \in C^2[x_0,x_1]$ with $g = 0$ near x_0 and x_1. Say that some $v \in \mathcal{U}$ minimizes the energy

$$\mathcal{E}_{\text{SecondDeriv}}(u) \stackrel{def}{=} \int_{x_0}^{x_1} \bigl(u''(x)\bigr)^2\,dx$$

over all $u \in \mathcal{U}$. If, in addition, $v \in C^4[x_0,x_1]$, then $v''''(x) = 0$ for all $x \in [x_0,x_1]$, i.e., v is a cubic polynomial over $[x_0,x_1]$.

Proof. If for some $y \in [x_0,x_1]$, $v''''(y) > 0$, we will obtain a contradiction; the case $v''''(y) < 0$ gives a similar contradiction.

So assume that for some $y \in [x_0,x_1]$, $v''''(y) > 0$. By continuity of v'''', we may assume $y \in (x_0,x_1)$. Let $\delta > 0$ be sufficiently small so that $x_0 < y-\delta$, $y+\delta < x_1$, and $v'''' > 0$ on all of $[y-\delta, y+\delta]$. Let

$$g(x) = (y-t-\delta)^4(y-t+\delta)^4$$

for $x \in (y-\delta, y+\delta)$, and $g(x) = 0$ otherwise. It follows that

$$\int_{x_0}^{x_1} \bigl(v''''(x)g(x)\bigr)\,dx > 0; \tag{7}$$

let us obtain a contradiction.

We easily see that $g \in C^2[x_0,x_1]$ and for any $\epsilon \in \mathbb{R}$ we have $v+\epsilon g \in \mathcal{U}$. Then for any $\epsilon \in \mathbb{R}$

$$\mathcal{E}_{\text{SecondDeriv}}(v+\epsilon g) = \int_{x_0}^{x_1} \bigl(v''(x)\bigr)^2\,dx + 2\epsilon \int_{x_0}^{x_1} \bigl(v''(x)g''(x)\bigr)\,dx + \epsilon^2 \int_{x_0}^{x_1} \bigl(g''(x)\bigr)^2\,dx.$$

By the minimizing property of v we have

$$\int_{x_0}^{x_1} \bigl(v''(x)\bigr)^2\,dx = \mathcal{E}_{\text{SecondDeriv}}(v) \le \mathcal{E}_{\text{SecondDeriv}}(v+\epsilon g),$$

hence for any $\epsilon \in \mathbb{R}$ we have

$$0 \ge 2\epsilon \int_{x_0}^{x_1} \bigl(v''(x)g''(x)\bigr)\,dx + \epsilon^2 \int_{x_0}^{x_1} \bigl(g''(x)\bigr)^2\,dx.$$

Taking $\epsilon \to 0$ it follows that

$$\int_{x_0}^{x_1} \bigl(v''(x)g''(x)\bigr)\,dx = 0.$$

Integrating by parts (and recalling that $g = 0$ near x_0 and x_1) we have

$$0 = \int_{x_0}^{x_1} \big(v''(x)g''(x)\big)\,dx = -\int_{x_0}^{x_1} \big(v'''(x)g'(x)\big)\,dx = \int_{x_0}^{x_1} \big(v''''(x)g(x)\big)\,dx.$$

But this contradicts (7). □

6. Some Subtle Points Regarding Energy Minimizers

Theorem 5.1 implies that if there is a $v \in \mathcal{U}_{\mathbf{t},\mathbf{y}}$ in (1) that minimizes the energy (2) over $\mathcal{U}_{\mathbf{t},\mathbf{y}}$, then v is cubic spline with break points $x_1, \ldots, x_{n-1}$ under the assumption that $v \in C^4[x_i, x_{i+1}]$ for $i = 0, \ldots, n-1$. A "variational" argument, similar to the proof of Theorem 5.1 shows that $v''(x_0) = v''(x_n) = 0$ (see the exercises). However, it is not at all clear that such a v should exist; in this section we explain that this existence is not a trivial matter.

6.1. **First Derivative Energies: Length, Dirichlet Integral, Etc.** Let us consider the following notations of "energy:"

(1)
$$\mathcal{E}_{\text{length}}(u) \stackrel{\text{def}}{=} \int_A^B \sqrt{1 + \big(u'(x)\big)^2}\,dx,$$
i.e., the length of the curve u from A to B, and

(2)
$$\mathcal{E}_{\text{Dirichlet}}(u) \stackrel{\text{def}}{=} \int_A^B \big(u'(x)\big)^2\,dx,$$
called the "Dirchlet integral."

Let us fix real numbers $A = x_0 < x_1 < \ldots < x_n = B$, a function $f\colon \mathbb{R} \to \mathbb{R}$, and a set

$$\hat{\mathcal{U}} = \hat{\mathcal{U}}_{f;\mathbf{t}} \stackrel{\text{def}}{=} \Big\{ u \in C^1[A,B] \;\Big|\; u(x_i) = f(x_i) \text{ for all } i \Big\}. \tag{8}$$

We have added a hat to our $\mathcal{U}$ to alert the reader that we only require that u have one continuous derivative over $[A, B]$, as opposed to the two derivatives required in $\mathcal{U}_{f;\mathbf{y}}$ of the form (3). The set $\hat{\mathcal{U}}$ is a reasonable analog of $\mathcal{U}$ when our energies involve only first derivatives; however, trying to minimize these energies over $\hat{\mathcal{U}} = \hat{\mathcal{U}}_{f;\mathbf{t}}$ turns out to be problematic, as we now explain.

Let us now argue—using geometry—that there is no $u \in \hat{\mathcal{U}}$ that minimizes $\mathcal{E}_{\text{length}}(u)$ over $\hat{\mathcal{U}}$ above. Consider the function $v = v(x)$ that connects the points $(x_i, f(x_i))$ by line segments (see Figure 11.2 of [A&G]); this is called the "piecewise linear interpolation" of f over $\mathbf{t}$, and also called a "bending line" or a "broken line" interpolation in Section 11.2 of [A&G]; note that v is not differentiable at each break point, and hence $v \notin \hat{\mathcal{U}}_{f;\mathbf{t}}$. However, since v has only finitely many discontinuities, it still makes sence to speak of $\mathcal{E}_{\text{length}}(v)$. We claim that for any $u \in \hat{\mathcal{U}}_{f;\mathbf{t}}$ we have

$$\mathcal{E}_{\text{length}}(v) < \mathcal{E}_{\text{length}}(u),$$

because the shortest line between two points is a straight line; however, for any $\epsilon > 0$ we can modify v a bit just a bit to the right of the break points $x_1, \ldots, x_{n-1}$ (using Hermite interpolation) to get a function $v_\epsilon \in \hat{\mathcal{U}}$ such that

$$\mathcal{E}_{\text{length}}(v_\epsilon) < \mathcal{E}_{\text{length}}(v) + \epsilon$$

(we'll draw a picture in class).

You can think of this as follows: the function $f(x) = x^2$ never attains a minimum (or maximum) value over the open interval $(13, 18)$ because the numbers 13 and 18 are "missing" from this interval; over the interval $[13, 18]$, the function $f(x) = x^2$ attains its minimum at $x = 13$ and its maximum $x = 18$.

By analogy, v above is "missing" from $\hat{\mathcal{U}}$, and we should somehow add in v. Of course, we can widen $\hat{\mathcal{U}}$ to include all any function that has a continuous derivative at all but finitely many points in $[A, B]$, but this is an "ad hoc" solution; in "real analysis" we learn much better ways to do this.

There is a "variational" argument similar to our proof of Theorem 5.1 that shows that $\mathcal{E}_{\text{Dirichlet}}(u)$ over $u \in \hat{\mathcal{U}}$ is similarly "minimized" by the same "bending line" v with the same remark that v is really "missing" from $\hat{\mathcal{U}}$. It turns out that the same is true (!) for the energy

$$\mathcal{E}_{\text{Strange}}(u) \stackrel{\text{def}}{=} \int_a^b e^{3f'(x)+17(f'(x))^{2020}} \sin\bigl(303 f'(x)\bigr)\, dx$$

using the same variational argument; note that this "energy" can be negative since it involves a sine.

6.2. **Energies Without Unique Minimizers and Weighted Energies.** Certain energy functions involving $u'(x)$ have many minimizers, including

$$\mathcal{E}_{\text{Area}}(u) \stackrel{\text{def}}{=} \int_A^B u'(x)\, dx, \quad \mathcal{E}_{\text{TotalVariation}}(u) \stackrel{\text{def}}{=} \int_A^B \bigl|u'(x)\bigr|\, dx.$$

Indeed, since

$$\int_A^B u'(x)\, dx = u(B) - u(A),$$

any $u \in \hat{\mathcal{U}} = \hat{\mathcal{U}}_{f;\mathbf{t}}$ (with $x_0 = A$ and $x_n = B$) has the same "area energy," $\mathcal{E}_{\text{Area}}(u)$. A similar observation holds for $\mathcal{E}_{\text{TotalVariation}}$, which we may briefly explain in class.

We point out that *weighted energies*, such as

$$\mathcal{E}_{\text{WeightedDirichlet}}(u) \stackrel{\text{def}}{=} \int_A^B \bigl(u'(x)\bigr)^2 w(x)\, dx$$

turn out to have energy "minimizers" (where the quotation marks are used to allow discontinuous first derivatives across the break points $x_1, \dots, x_{n-1}$), but these minimizers depend on the "weight function" $w(x)$. These minimizers are only piecewise polynomials when $w(x)$ is a constant in each interval (x_i, x_{i+1}).

6.3. **Closure, Real Analysis, and Energy Minimizing.** Let us briefly sketch how one minimizes energy using "closure" and "real analysis." This discussion wil NOT be part of CPSC 303 this term.

To minimize $\mathcal{E}(u) = \mathcal{E}_{\text{length}}(u)$ or $\mathcal{E}(u) = \mathcal{E}_{\text{Dirichlet}}(u)$ over $u \in \hat{\mathcal{U}}_{f;\mathbf{t}}$ as above, we choose a sequence $u_1, u_2, \dots \in \hat{\mathcal{U}}_{f;\mathbf{t}}$ such that as $i \to \infty$, the limit of $\mathcal{E}(u_i)$ exists and is the smallest possible such limit. One then proves that there is a reasonable of a "limiting function"

$$\lim_{i\to\infty} u_i;$$

this function turns out to be the broken line described earlier.

Here is another way to think of things.

We try to think of a reasonable notation of "distance" between two functions in $\hat{\mathcal{U}}_{f;\mathbf{t}}$. When we are interested in $\mathcal{E}_{\text{length}}(u)$ and $\mathcal{E}_{\text{Dirichlet}}(u)$, there is a very good notion of distance to use, namely

$$\rho(u,w) = \rho_{\text{Dirichlet}}(u,w) \stackrel{\text{def}}{=} \sqrt{\int_A^B \big(u'(x) - w'(x)\big)^2\, dx};$$

it takes some work to verify that this $\rho_{\text{Dirichlet}}$ satisfies the usual properties we require of distances, but this a standard part of "real analysis;" for example, since $u(x_0) = w(x_0) = y_0$ and $u(x_1) = w(x_1) = y_1$ we have that $\rho(u,w) = 0$ iff $u = w$.

Once we have a notion of "distance" between elements of a set, there is a standard way to define the "closure" of the set. It turns out that the closure, W, of $\hat{\mathcal{U}}_{f;\mathbf{c}}$ under $\rho_{\text{Dirichlet}}$,

(1) the "bending line," v, of interest to us, lies in W;
(2) W is a standard type of very nice vector space (often called *Sobolev spaces*, which involve "weak derivatives") and can be described in a number of different, convenient ways;
(3) $\mathcal{E}_{\text{length}}$ and $\mathcal{E}_{\text{Dirichlet}}$ are *continuous* functions on W with respect to $\rho_{\text{Dirichlet}}$; and
(4) any function $W \to \mathbb{R}$ that is continuous with respect $\rho_{\text{Dirichlet}}$ and bounded from below attains a minimum value somewhere on W.

So just as the real interval $(13, 18)$ adds the "missing" points 13 and 18 when we take the closure of this interval under the usual distance $\rho(x,y) = |x - y|$, our missing "bending lines" are added when we the take closure of $\hat{\mathcal{U}}_{f;\mathbf{t}}$ (8) under a good metric.

6.4. **Back to Second Derivative Energy.** When we consider the energy (2), based on the square of the second derivative, a good notation of distance is

$$\rho(u,w) = \rho_{\text{SecondDerivEnergy}}(u,w) \stackrel{\text{def}}{=} \sqrt{\int_A^B \big(u''(x) - w''(x)\big)^2\, dx}.$$

It turns out that the minimizer, v, of this energy will automatically (meaning by standard results) have a "nice" first derivative, v', but it is not clear that v'' will exist at the break points; this can be proven by a variational argument (see the exercises).

Exercises

(1) Here will appear Exercise 1.
(2) Here will appear Exercise 2.
(3) Here will appear Exercise 3.
(4) Etc.

Department of Computer Science, University of British Columbia, Vancouver, BC V6T 1Z4, CANADA.

E-mail address: jf@cs.ubc.ca

URL: http://www.cs.ubc.ca/~jf