An Expansion of the Derivation of the Spline Smoothing Theory

Alan Kaylor Cline

The classic paper "Smoothing by Spline Functions", *Numerische Mathematik* **10,** 177-183 (1967) by Christian Reinsch showed that natural cubic splines were the solutions to a novel formulation of the data smoothing problem. Reinsch employed what he called "standard methods of the calculus of variations" to obtain his results and this included the use of Euler-Lagrange equations. It is my intent here to expand that derivation and to avoid any reference to the calculus of variations. The material here simply replaces the first paragraph of Section 3 of Reinsch's paper. I have tried to present it so that only calculus and linear algebra are required for understanding.

1. The Problem

Consider that we are given a set of triples $\{(x_i, y_i, \delta y_i)\}_{i=1}^n$ and a nonnegative value S such that $x_1 < x_2 < \cdots < x_n$ and $\delta y_i > 0$ for $i = 1,2,...,n.$ We seek to find a function f to minimize

$$\int_{x_1}^{x_n} g''(x)^2 dx$$

over all functions g with two continuous derivatives on $[x_1, x_n]$ such that

$$\sum_{i=1}^{n} \left(\frac{g(x_i) - y_i}{\delta y_i} \right)^2 \le S.$$

In the very terse first paragraph of Section 3, Reinsch showed:

1. Any solution to the problem must be a natural cubic spline with knots $\{x_i\}_{i=1}^n$,

2. For $S > 0$, the gaps in the third derivatives of a solution are proportional to the weighted residuals in the approximations. That is for some non-negative value of p

$$f'''(x_i)_- - f'''(x_i)_+ = 2p \frac{f(x_i) - y_i}{\delta y_i^2}$$

for $i = 1,2,...,n,$ where the subscripts + and - indicate right- and left-sided derivatives, respectively, and the interpretation at the endpoints is that $f'''(x_1)_- = 0$ and $f'''(x_n)_+ = 0$.

2. The Expansion

We begin with a lemma that relates integrals of natural cubic splines to their third derivative gaps.

Lemma 1: If g is a natural cubic spline with knots $x_1 < x_2 < \cdots < x_n$ and l is a function with two continuous derivatives on $[x_1, x_n]$ then

$$\int_{x_1}^{x_n} g''(x) l''(x) dx = \sum_{i=1}^{n} l(x_i)(g'''(x_i)_+ - g'''(x_i)_-),$$

where $g'''(x_1)_- = 0$ and $g'''(x_n)_+ = 0$.

Proof: Although g''' is not continuous on $[x_1, x_n]$, it is continuous on each interval (x_i, x_{i+1}), for $i = 1,2,...,n,$. Integrating by parts on each of the smaller intervals we find

$$\int_{x_1}^{x_n} g''(x)l''(x)dx = \sum_{i=1}^{n-1}\left(g''(x_{i+1})l'(x_{i+1}) - g''(x_i)l'(x_i) - \int_{x_i}^{x_{i+1}} g'''(x)l'(x)dx \right).$$

However the terms within the summation can be separated and all of the products $g''(x_i)l'(x_i)$ (except the first and last) will cancel each other resulting in

$$= g''(x_n)l'(x_n) - g''(x_1)l'(x_1) - \sum_{i=1}^{n-1}\left(\int_{x_i}^{x_{i+1}} g'''(x)l'(x)dx \right).$$

But from the natural end-conditions of g and the fact that g''' is piece-wise constant on each of the intervals (x_i, x_{i+1}), we have

$$= 0 - \sum_{i=1}^{n-1} g'''(x_i)_+ (l(x_{i+1}) - l(x_i))$$

$$= -\sum_{i=1}^{n-1} g'''(x_i)_+ l(x_{i+1}) + \sum_{i=1}^{n-1} g'''(x_i)_+ l(x_i)$$

$$= -\sum_{i=2}^{n} g'''(x_{i-1})_+ l(x_i) + \sum_{i=1}^{n-1} g'''(x_i)_+ l(x_i).$$

Finally by noticing that $g'''(x_{i-1})_+ = g'''(x_i)_-$, we obtain

$$= -\sum_{i=2}^{n} g'''(x_i)_- l(x_i) + \sum_{i=1}^{n-1} g'''(x_i)_+ l(x_i)$$

$$= \sum_{i=1}^{n} (g'''(x_i)_+ - g'''(x_i)_-)l(x_i).$$

□

Reinsch's first proposition is an easy consequence of Lemma 1.

Any solution to the problem must be a natural cubic spline with knots $\{x_i\}_{i=1}^n$

Proof: If f is a solution to the problem and g is a natural cubic spline that interpolates f on the set $\{x_i\}_{i=1}^n$ (i.e., $g(x_i) = f(x_i)$ for $i = 1,2,...,n$), then g certainly satisfys

$$\sum_{i=1}^{n}\left(\frac{g(x_i) - y_i}{\delta y_i} \right)^2 \le S.$$

since f does. Furthermore, expressing f as $(f - g) + g$, we have

$$\int_{x_1}^{x_n} f''(x)^2 dx = \int_{x_1}^{x_n} (f''(x) - g''(x))^2 dx + \int_{x_1}^{x_n} (f''(x) - g''(x))g''(x)dx + \int_{x_1}^{x_n} g''(x)^2 dx.$$

But from Lemma 1 and the interpolation conditions, we have

$$= \int_{x_1}^{x_n} (f''(x) - g''(x))^2 dx + 2\sum_{i=1}^{n} (f(x_i) - g(x_i))(g'''(x_i)_+ - g'''(x_i)_-) + \int_{x_1}^{x_n} g''(x)^2 dx$$

$$= \int_{x_1}^{x_n} (f''(x) - g''(x))^2 dx + 0 + \int_{x_1}^{x_n} g''(x)^2 dx.$$

The first term is clearly non-negative, and thus

$$\int_{x_1}^{x_n} f''(x)^2 dx \geq \int_{x_1}^{x_n} g''(x)^2 dx.$$

Furthermore, the inequality is strict (and thus the solution *must* be a natural cubic spline) unless

$$\int_{x_1}^{x_n} (f''(x) - g''(x))^2 dx = 0.$$

However for this to happen $f'' - g'' \equiv 0$ on $[x_1, x_n]$. Integrating twice, we have that $f - g$ must be a linear polynomial on $[x_1, x_n]$, yet from the interpolation conditions, that linear must be identically zero. Thus f cannot be a solution to the problem without being a natural cubic spline with knots $\{x_i\}_{i=1}^n$. □

The next lemma simply says that unless two vectors have exactly the same direction then we can find a vector with a positive component in direction of one but a negative component with respect to the other.

Lemma 2. Let u and v be non-zero *n*-vectors. If for no non-negative scalar p does $u = pv$, then there exists an *n*-vector w so that $w^T u > 0$ and $w^T v < 0$.

Proof: Let $\| z \|$ denote the euclidean norm $(z^T z)^{1/2}$. From the Cauchy-Schwartz inequality, since u and v are not in the same direction,

$$u^T v < \|u\| \|v\|.$$

Thus,

$$\frac{u^T v}{\|v\|} < \|u\|$$

and

$$0 < \|u\| - \frac{u^T v}{\|v\|}.$$

Similarly

$$0 < \|v\| - \frac{u^T v}{\|u\|}.$$

Let $w = \frac{u}{\|u\|} - \frac{v}{\|v\|}$, then

$$w^T u = \frac{u^T u}{\|u\|} - \frac{v^T u}{\|v\|}$$
$$= \|u\| - \frac{u^T v}{\|v\|}$$
$$> 0.$$

Yet

$$w^T v = \frac{u^T v}{\|u\|} - \frac{v^T v}{\|v\|}$$
$$= \frac{u^T v}{\|u\|} - \|v\|$$
$$< 0.$$

□

The vector w will be used to show that unless a particular function is a solution to our problem, slight perturbations can be made that reduce the objective function $\int_{x_1}^{x_n} g''(x)^2 dx$ without violating the constraints $\sum_{i=1}^{n} \left(\frac{g(x_i) - y_i}{\delta y_i} \right)^2 \leq S.$

Reinsch's second proposition is

If $S > 0$, *for some non-negative value of* p *then for* $i = 1, 2, ..., n$,

$$f'''(x_i)_- - f'''(x_i)_+ = 2p \frac{f(x_i) - y_i}{\delta y_i^2}.$$

Proof: Suppose for some natural cubic spline g with knots $\{x_i\}_{i=1}^n$ there is no non-negative value of p for which

$$g'''(x_i)_- - g'''(x_i)_+ = 2p \frac{g(x_i) - y_i}{\delta y_i^2}.$$

for $i = 1, 2, ..., n$. In particular, this implies that the quantities $g'''(x_i)_- - g'''(x_i)_+$ are not all zero. Either the quantities $\frac{g(x_i) - y_i}{\delta y_i^2}$ are also not all zero or they are all zero. Suppose initially that they are not all zero. Lemma 2 then applies and we can find values w_i for $i = 1, 2, ..., n$ so that

$$\sum_{i=1}^{n} w_i (g'''(x_i)_- - g'''(x_i)_+) > 0$$

and

$$\sum_{i=1}^{n} w_i \frac{g(x_i) - y_i}{\delta y_i^2} < 0.$$

If we let l be any function with two continuous derivatives on $[x_1, x_n]$ that interpolates the data pairs $\{(x_i, w_i)\}_{i=1}^n$ (i.e., $l(x_i) = w_i$, for $i = 1,2,...,n$), then

$$\sum_{i=1}^{n} l(x_i)(g'''(x_i)_- - g'''(x_i)_+) > 0$$

and

$$\sum_{i=1}^{n} l(x_i)\frac{g(x_i) - y_i}{\delta y_i^2} < 0.$$

Now for any scalar λ

$$\begin{aligned}\int_{x_1}^{x_n} (g''(x) + \lambda l''(x))^2 dx &= \int_{x_1}^{x_n} g''(x)^2 dx + 2\lambda \int_{x_1}^{x_n} g''(x)l''(x)dx + \lambda^2 \int_{x_1}^{x_n} l''(x)^2 dx \\ &= \int_{x_1}^{x_n} g''(x)^2 dx + 2\lambda \sum_{i=1}^{n} l(x_i)(g'''(x_i)_+ - g'''(x_i)_-) + \lambda^2 \int_{x_1}^{x_n} l''(x)^2 dx \\ &= \int_{x_1}^{x_n} g''(x)^2 dx - 2\lambda \sum_{i=1}^{n} l(x_i)(g'''(x_i)_- - g'''(x_i)_+) + \lambda^2 \int_{x_1}^{x_n} l''(x)^2 dx.\end{aligned}$$

Notice that for all positive values of λ the middle term on the right-hand side is negative and, for sufficiently small positive values of λ, the sum of the last two terms is negative. Thus for those values of λ

$$\int_{x_1}^{x_n} (g''(x) + \lambda l''(x))^2 dx < \int_{x_1}^{x_n} g''(x)^2 dx\,.$$

However, at the same time,

$$\sum_{i=1}^{n}\left(\frac{g(x_i) + \lambda l(x_i) - y_i}{\delta y_i^2}\right)^2 = \sum_{i=1}^{n}\left(\frac{g(x_i) - y_i}{\delta y_i^2}\right)^2 + 2\lambda \sum_{i=1}^{n} l(x_i)\left(\frac{g(x_i) - y_i}{\delta y_i^2}\right) + \lambda^2 \sum_{i=1}^{n}\left(\frac{l(x_i)}{\delta y_i^2}\right)^2$$

Again for all positive values of λ the middle term on the right-hand side is negative and, for sufficiently small positive values of λ, the sum of the last two terms is negative. If

$$\sum_{i=1}^{n}\left(\frac{g(x_i) - y_i}{\delta y_i}\right)^2 \le S,$$

then for sufficiently small positive values of λ

$$\sum_{i=1}^{n}\left(\frac{g(x_i) + \lambda l(x_i) - y_i}{\delta y_i^2}\right)^2 < S.$$

The conclusion is that, if g does not satisfy the proportionality relation of the hypothesis, then a function $g + \lambda l$ satisfies the constraints and reduces the objective function. Thus g could not have been a solution.

Alternatively, we assume that all of the quantities $\frac{g(x_i) - y_i}{\delta y_i^2}$ are zero. It is still possible to find a vector w so that

$$\sum_{i=1}^{n} w_i (g'''(x_i)_- - g'''(x_i)_+) > 0.$$

As above, a function l could be found for which

$$\int_{x_1}^{x_n} (g''(x) + \lambda l''(x))^2 dx < \int_{x_1}^{x_n} g''(x)^2 dx$$

for sufficiently small positive values of λ. However since $S > 0$, for small values of λ

$$\sum_{i=1}^{n} \left(\frac{g(x_i) + \lambda l(x_i) - y_i}{\delta y_i^2} \right)^2 = \lambda^2 \sum_{i=1}^{n} \left(\frac{l(x_i)}{\delta y_i^2} \right)^2 \le S.$$

So once again, g could not be the solution. □

The final part of this expansion of Reinsch's derivation shows that, if at the solution .

$$\sum_{i=1}^{n} \left(\frac{f(x_i) - y_i}{\delta y_i} \right)^2 < S,$$

then

$$f'''(x_i)_- - f'''(x_i)_+ = 0,$$

for $i = 1, 2, ..., n$, (i.e., the parameter $p = 0$), since otherwise the argument could be repeated to produce a better solution. Of course, these conditions result in no gaps in the third derivatives. A solution must then be a single cubic. However the natural end conditions then imply that the cubic solution is actually a linear. The result is that either a solution rests at the edge of the constraints

$$\sum_{i=1}^{n} \left(\frac{f(x_i) - y_i}{\delta y_i} \right)^2 = S$$

or is a straight line.

Additional Note on Reinsch's Paper:

At the end of Section 4, Reinsch's presents an iterative algorithm and then a final paragraph where he comments that one could also apply a Newton's method to find the reciprocal of p. With that, he claims, "When tested with few examples, convergence was always reached with a slightly reduced number of iterations.". Although he does not state it explicitly, the ALGOL code in Section 5, actually implements the iteration to find the reciprocal of p.

As Reinsch states in the second paper, "Smoothing by Spline Functions II", *Numerische Mathematik* **16** 451-454 (1971), the iteration to find the reciprocal of p may not converge. There he suggests that instead of using Newton's method on the equation $F(p) = S^{1/2}$, that faster (and guaranteed) convergence is obtained with the equation $1/F(p) = S^{-1/2}$. The only change that this would require from the algorithm in Section 4 of the first paper, is the replacement of the Newton update

$$p \leftarrow p + (e - (Se)^{1/2})/(f - p \times g)$$

by

$$p \leftarrow p + e((e/S)^{1/2} - 1)/(f - p \times g).$$