Construction of monotone smoothing splines using a branch and bound method

Beräkning av monotona utjämnande splines med en branch and bound-algoritm

Malte Larsson

Master's thesis
2020:E38

Lund Institute of Technology
Lund University

Faculty of Engineering
Centre for Mathematical Sciences
Mathematics

Centrum Scientiarum Mathematicarum

Construction of monotone smoothing splines using a branch and bound method

Malte Larsson

Abstract

The purpose of this report is to investigate a branch and bound algorithm designed to construct monotone smoothing splines. The splines are defined as the solution of a minimization problem, where one wants to find a curve with non-negative derivative everywhere, that fits a data set consisting of points in the plane, while also being sufficiently smooth. The report describes the theory necessary to understand the mathematical background of the algorithm. The theory shows existence and uniqueness of a minimizer and uses the Karush-Kuhn-Tucker (KKT) conditions for vector spaces to reduce the problem to a finite dimensional optimization problem. The report also shows some results that can be used to improve the algorithm and reduce the search space from 2^n to 1.84^n, where n is the number of data points. In practice the running time is often faster, although it can depend on some parameters as well as the data set.

Acknowledgements

First and foremost, I would like to thank my supervisor Sara Maad Sasane. She was the one who suggested this project to me and has been of great help during my work on this master thesis. I also want to thank my friends and family for all their support throughout my entire education.

Contents

1 Introduction

This report will explore an algorithm for constructing monotone smoothing splines. The purpose of splines is to fit a curve to a set of data points (t_k, α_k), on an interval $[0, T]$. Fitting a curve to a set of data points can be done in several ways. One way is to fit a parameterized function to the data. Then the parameters could be estimated using for example least squares regression. Another approach is to use smoothing splines. Then the approach is to find a curve that fits the given data, with the additional condition that the curve shouldn't change too fast. More precisely, we want to find the minimizer of

$$\int_0^T (\ddot{x}(t))^2 dt + \lambda \sum_{k=0}^{n} (x(t_k) - \alpha_k)^2.$$

Monotonous smoothing splines satisfy the additional condition that $\dot{x}(t) \geq 0$, at all points on the interval. We will also include some additional constraints, such that the function should be positive and bounded by some maximal value.

Previous work on monotonous smoothing splines can be found for example in Egerstedt and Martin [3]. The algorithm that is explored in this report is a branch and bound method outlined in Maad Sasane [1]. The idea is that on each interval between two data points, the function must be of one of two types. The next chapter will include large parts of the mathematical theory and proofs necessary to show this. Chapter 3 will describe the actual algorithm, and give some proofs of correctness of it. Here is also shown some improvements that can be made to reduce the algorithm's time complexity. In Chapter 4 we discuss the performance of the algorithm.

2 Monotone smoothing splines

In this section we find some mathematical properties that is needed to construct an algorithm that can find solutions to the problem of computing monotone smoothing splines. The general idea is to show that on each interval between the spline knots, there are two possibilities for what type of function there can be. If it is known what values the function and its derivative should have at the spline knots, then one can find exactly what the optimal function should be. For this reason, the problem can be rephrased as an optimization problem over these values at the spline knots.

First of all, let us formulate the problem again with a few more details. Given a set of points (t_i, α_i), $i = 1, 2, \ldots, n$, with $t_1 = 0, t_n = T$, we want to find the minimizer of

$$\int_0^T \ddot{x}(t)^2 dt + \lambda \sum_{i=1}^{n} (x(t_i) - \alpha_i)^2 \tag{1}$$

subject to

$$\begin{cases} x \in H^2(0,T) \\ \dot{x} \geq 0 \\ x(0) = 0 \\ x(T) \leq x_{max}. \end{cases}$$

Here $H^2(0,T)$ is the Sobolov space of twice weakly differentiable functions, defined on the interval $(0,T)$, such that the L^2-norm is finite for the function, and its two weak derivatives. A function u defined on an interval $[a,b]$ has a weak derivative v if

$$\int_a^b u(t)\varphi'(t)dt = -\int_a^b v(t)\varphi(t)dt,$$

for all smooth functions $\varphi(x)$ with $\varphi(a) = \varphi(b) = 0$. This definition essentially says that partial integration should work as expected. When a function is differentiable, the derivative is the same as the weak derivative. One example of functions that are relevant for this problem that are weakly differentiable but not differentiable, are functions with left and right derivatives, that do not coincide at a finite number of points.

This space is a Hilbert space which is one reason for using this space rather than C^2, the space of twice continuously differentiable functions. The Sobolev space has the norm

$$||x||_{H^2} = ||x||_{L^2} + ||\dot{x}||_{L^2} + ||\ddot{x}||_{L^2} = \left(\int_0^T x^2 dt\right)^{1/2} + \left(\int_0^T \dot{x}^2 dt\right)^{1/2} + \left(\int_0^T \ddot{x}^2 dt\right)^{1/2}.$$

One useful property is that $H^2(0,T) \subset C^1(0,T)$, i.e. the functions are continuously differentiable, and there exists a constant C such that $\sup_{t\in[0,t]} |x(t)| \leq C||x||_{H^2}$, see e.g. Renardy and Rogers [8].

In the introduction, we mentioned that we want the function to be bounded by some value x_{max}. Since the function is increasing it is enough to impose this condition at $t = T$. The condition $x(0) = 0$, could also be replaced by $x(0) \geq 0$, depending on the situation, the proofs below are valid in both cases. This would for example be the case if one wanted to estimate a cumulative distribution function given data. Then one could also use $x_{max} = 1$. The left endpoint 0 of the entire interval is just chosen for convenience, as the data could be shifted in t-direction to an arbitrary starting point.

The main difficulty with this problem is how to deal with the monotonicity criterion. We will therefore first show what happens when this criterion is not present, which also yields a result that will be useful later.

Theorem 1. *Let $x(t) \in H^2(0,1)$ with $x(0) = x_0$, $x(1) = x_1$, $\dot{x}(0) = \dot{x}_0$ and $\dot{x}(1) = \dot{x}_1$. Then the minimizer of*

$$\int_0^1 \ddot{x}(t)^2 dt \tag{2}$$

will be a third degree polynomial.

Proof. Let $x(t)$ be a third degree polynomial with correct values for $x(0)$, $x(1)$, $\dot{x}(0)$ and $\dot{x}(1)$. Any other valid function can be written as $x + h$ where $h(t)$ is an H^2 function with

$$h(0) = h(1) = \dot{h}(0) = \dot{h}(1) = 0. \tag{3}$$

Then

$$\int_0^1 (\ddot{x} + \ddot{h})^2 dt = \int_0^1 (\ddot{x}^2 + 2\ddot{x}\ddot{h} + \ddot{h}^2) dt.$$

Since x is a third degree polynomial, $\ddot{x} = At + B$ and we get

$$\int_0^1 2\ddot{x}\ddot{h} dt = 2\int_0^1 (At\ddot{h} + B\ddot{h}) dt = 2A\int_0^1 t\ddot{h} dt + 2B\int_0^1 \ddot{h} dt.$$

The second term becomes $\dot{h}(1) - \dot{h}(0) = 0$ and with integration by parts we get

$$\int_0^1 t\ddot{h} dt = [t\ddot{h}]_0^1 - \int \ddot{h} dt = (1 \cdot \dot{h}(1) - 0 \cdot \dot{h}(0)) - (h(1) - h(0)) = 0.$$

This shows that

$$\int_0^1 2\ddot{x}\ddot{h} dt = 0.$$

Together with $\int_0^1 \ddot{h}^2 \geq 0$ we get

$$\int_0^1 (\ddot{x} + \ddot{h})^2 dt = \int_0^1 (\ddot{x}^2 + 2\ddot{x}\ddot{h} + \ddot{h}^2) dt \geq \int_0^1 \ddot{x}^2 dt,$$

and thus the third degree polynomial x is optimal. In addition we only get equality if $\int_0^1 \ddot{h}^2 dt = 0$ and together with (3) we see that this can only occur if $h \equiv 0$. Therefore our minimizer is unique.

□

If the values of the function and its derivatives are known at the spline knots, then the sum in (1) is fixed and on every subinterval a third degree polynomial is optimal. This would be a valid function since the function is continuously differentiable everywhere, and is twice differentiable everywhere except at the spline knots. If we consider these values to be variables, we could minimize the expression depending on these variables.

The solution will depend on the choice of λ. There are ways to estimate the best choice of λ, for example by leaving out some points and fitting the curve to the rest of the points and see what choice of λ gives best fit to the removed points. More on this can be found in Wang [2]. We will not discuss further the problem of finding the optimal choice of λ.

We will now turn our attention back to the problem of finding monotone smoothing splines. The rest of this section will first of all show that there is a unique minimizer to (1). Once we know that this exists, we will use an infinite dimensional version of the Karush-Kuhn-Tucker (KKT) conditions, to find some properties of this minimizer. Finally this can be used to characterize the optimal function on an interval when the third degree polynomial doesn't work.

2.1 Existence and uniqueness of the minimizer

To show the existence of the minimizer we will utilize a theorem that is stated below. First however we will define some additional concepts that are used in this theorem.

Some concepts from functional analysis is used. One of them is the concept of dual spaces. Given a normed vector space X, we can define a space consisting of all linear and bounded functionals, going from X to $\mathbb{R}$. A functional f is bounded if there exists a constant C such that $|f(x)| \leq C||x||_X$ for all $x \in X$. We denote this space by X^*. When applying a function $x^* \in X^*$ on an element $x \in X$ the notation $\langle x, x^* \rangle$ is often used. A space is called reflexive if the dual space of the dual space can be identified with the original space. This is always true for Hilbert spaces, see e.g. [8].

Definition 2.1. Let $\{s_n\}$ be a sequence of real numbers. Let E be the set of numbers x in $\mathbb{R} \cup \{-\infty, +\infty\}$, such that $s_{n_k} \to x$ for some subsequnce $\{s_{n_k}\}$. We define the limit inferior and the limit superior

$$\liminf_{n\to\infty} s_n = \inf E \quad \text{and} \quad \limsup_{n\to\infty} s_n = \sup E.$$

Whenever a sequence converges we have that $\limsup s_n = \liminf s_n = \lim s_n$. However, while $\lim s_n$ doesn't have to exist, $\liminf_{n\to\infty} s_n$ and $\limsup_{n\to\infty} s_n$ always exist when $-\infty$ and ∞ are amended to the real numbers. This can be seen as the set E will never be empty. If the sequence is unbounded it will contain either ∞ or $-\infty$, otherwise it is part of a compact set and will have a convergent subsequence.

Definition 2.2. Let H be a Hilbert space. A sequence $x_n \in H$ is weakly convergent to $x \in H$ if

$$f(x_n) \to f(x)$$

for all $f \in H^*$. We denote this by $x_n \rightharpoonup x$.

Similarly to how a set can be closed if all convergent sequences converge to a point in the set we say that a set is weakly closed if $x_n \rightharpoonup x$ implies that x belongs to the set.

The following theorem from Struwe [7] is used to show the existance of a minimizer.

Theorem 2. *Suppose V is a reflexive Banach space with norm $||\cdot||$, and let $M \subset V$ be a weakly closed subset of V. Suppose $E : M \to \mathbb{R} \cup \{\infty\}$ is a coercive and (sequentially) weakly lower semi-continuous on M with respect to V, that is the following conditions are fulfilled*

1. *$E(u) \to \infty$ as $||u|| \to \infty, u \in M$*
2. *For any $u \in M$, any sequence (u_n) in M such that $u_n \rightharpoonup u$ in V there holds:*

$$E(u) \leq \liminf_{n\to\infty} E(u_n).$$

Then E is bounded from below on M and attains its infimum in M.

We use this theorem to show the existence of a solution to our minimization problem. In order to use it we will also need the following lemma to show that the feasible set is weakly closed.

Theorem 3 (Mazur's lemma)**.** *Let X be a Banach space and suppose*

$$u_n \rightharpoonup u$$

in X. Then there exists a function $N : \mathbb{N} \mapsto \mathbb{N}$, and a sequence of sets of real numbers $\{\alpha(n)_k\}_{k=n}^{N(n)}$ such that $\alpha(n)_k \geq 0$ and $\sum_{k=n}^{N(n)} \alpha(n)_k = 1$ such that the sequence

$$v_n := \sum_{k=n}^{N(n)} \alpha(n)_k u_k$$

converges strongly to u in X.

Essentially the theorem states that given a weakly convergent sequence, one can construct a new sequence consisting of convex combinations of the elements in the first sequence, such that the new sequence converges strongly to the limit element.

Lemma 1. *There exists a constant C such that for $x(t) \in H^2((0,T))$ with $x(0) \geq 0$, $\bar{x} = \frac{1}{T}\int_0^T x dt$ and $\bar{\dot{x}} = \frac{1}{T}\int_0^T \dot{x} dt$, we have*

$$||x||_{H^2} \leq C(||\ddot{x}||_{L^2} + |\bar{x}| + |\bar{\dot{x}}|).$$

Proof. Without loss of generality, assume that the interval is $(0,1)$. By removing the mean value of the function we get according to the Poincaré inequality, see e.g. Evans [10], that there exists a positive constant C, independent of x, such that

$$||x - \bar{x}||_{L^2} = \left(\int_0^1 (x-\bar{x})^2 dt\right)^{1/2} \leq C\left(\int_0^1 \dot{x}^2 dt\right)^{1/2} = C||\dot{x}||_{L^2}.$$

Using the triangle inequality, we get

$$||x||_{L^2} \leq ||x - \bar{x}||_{L^2} + ||\bar{x}||_{L^2} \leq C||\dot{x}||_{L^2} + |\bar{x}|$$

Similarly

$$||\dot{x} - \bar{\dot{x}}||_{L^2} \leq C||\ddot{x}||_{L^2}.$$

and

$$||\dot{x}||_{L^2} \leq C||\ddot{x}||_{L^2} + |\bar{\dot{x}}|.$$

We now get

$$\begin{aligned} ||x||_{H^2} = ||x||_{L^2} + ||\dot{x}||_{L^2} + ||\ddot{x}||_{L^2} \leq (C+1)||\dot{x}||_{L^2} + ||\ddot{x}||_{L^2} + |\bar{x}| \leq \\ \leq (C^2 + C + 1)||\ddot{x}||_{L^2} + |\bar{x}| + (C+1)|\bar{\dot{x}}|, \end{aligned}$$

which finishes the proof.

□

Theorem 4. *Let* $X = \{x \in H^2((0,T)); x(0) = 0, \dot{x}(t) \geq 0 \, \forall t \in (0,T), x(T) \leq x_{max}\}$, *with* $m \geq 2$. *Then there exists a unique solution* x^* *to the problem*

$$\min_{x \in X} \left(\frac{1}{2} \int_0^T \ddot{x}(t)^2 dt + \frac{1}{2} \sum_{i=1}^{m} (x(t_i) - \alpha_i)^2 \right). \tag{4}$$

Proof. We apply Theorem 2 to the problem to show that the problem has a unique minimizer. First of all $X \subset H^2(0,T)$ which is reflexive. We also need to show that our space X is weakly closed. Suppose that u_n converges weakly to u_0. Then Mazur's lemma shows that there exists a sequence v_n that converges strongly to u_0, where each element in (v_n) is a convex combination of elements in (u_n). Since X is convex it follows that $(v_n) \in X$ and since X is closed it follows that u_0 must lie in X. Thus X is weakly closed.

We start by showing the coerciveness. Above we showed that

$$||x||_{H^2} \leq C||\ddot{x}||_{L^2} + |\bar{x}| + |\bar{\dot{x}}|.$$

Thus for $||x||_{H^2}$ to go to infinity, either $||\ddot{x}||_{L^2}$, $|\bar{x}|$ or $|\bar{\dot{x}}|$ has to go to infinity. Neither of the last two terms can go to infinity. Since the function is bounded the maximal value of $|\bar{x}|$ is $T \cdot x_{max}$. If $|\bar{\dot{x}}| \to \infty$ then

$$\int_0^T \dot{x} dt = x(T) - x(0) \to \infty.$$

Since $x(0) \geq 0$, $x(T) \to \infty$ which contradicts that the function is bounded. If $||\ddot{x}||_{L^2} \to \infty$ then the integral part of (4) goes to infinity, and since the sum is non-negative, the entire functional goes to infinity.

Now we show the weak lower semi-continuity. Define g by $g\colon x \mapsto \int_0^T \ddot{x}\ddot{x}_0 dt$ and note that $g \in X^*$. It is clearly linear and that it is bounded follows from Cauchy-Schwarz inequality

$$||g(x)|| = \int_0^T \ddot{x}\ddot{x}_0 dt \leq ||\ddot{x}||_{L^2}||\ddot{x}_0||_{L^2} \leq ||\ddot{x}_0||_{L^2}||x||_{H^2}.$$

Suppose that $x_n \rightharpoonup x_0$. We have that

$$0 \leq \int_0^T (\ddot{x}_n - \ddot{x}_0)^2 dt$$

Therefore

$$0 \leq \liminf_{n\to\infty} \int_0^T (\ddot{x}_n - \ddot{x}_0)^2 dt = \liminf_{n\to\infty} \int_0^T \ddot{x}_n^2 dt - 2\int_0^T \ddot{x}_0\ddot{x}_n dt + \int_0^T \ddot{x}_0^2 dt$$

Since $g \in X^*$ and x_n converges weakly to x_0 we get that the middle term converges to $2\int_0^T \ddot{x}_0^2 dt$. Thus

$$0 \leq \liminf_{n\to\infty}(\int_0^T \ddot{x}_n^2 dt - \int_0^T \ddot{x}_0^2 dt)$$

The last term does not depend on n and can thus be moved out of the limit and to the other side of the inequality. This gives us

$$\int_0^T \ddot{x}_0^2 dt \leq \liminf \int_0^T \ddot{x}_n^2 dt,$$

which is exactly the lower semicontinuity for the first term.

Now we will prove this for the sum as well. For fixed $t \in [0,T]$, let $h_t(x) = x(t)$. This is a linear functional and it is bounded as well since

$$|h_t(x)| = |x(t)| \leq \sup_{t\in[0,T]} |x(t)| \leq C||x||_{H^2}.$$

Thus if $x_k \rightharpoonup x$ we get $h_t(x_k) \to h_t(x)$, or equivalently $x_k(t) \to x(t)$. Due to continuity of the square function we get that

$$\lim_{k\to\infty} \sum_{i=1}^m (x_k(t_i) - \alpha_i)^2 = \sum_{i=1}^m (x(t_i) - \alpha_i)^2$$

We will prove the uniqueness of the minimizer by showing that the function we minimize in (4) is strictly convex. We denote this function by f. We start by showing that $x \mapsto \int_0^T x^2 dt$ is strictly convex. Let $u, v \in X$, and λ, μ with $0 \leq \lambda, \mu \leq 1$ and $\lambda + \mu = 1$. Then

$$\begin{aligned}\int_0^T (\lambda u + \mu v)^2 dt = \lambda^2 \int_0^T u^2 dt + \mu^2 \int_0^T v^2 dt + 2\lambda\mu \int_0^T uv dt \leq \\ \leq \lambda^2 ||u||_{L^2}^2 + \mu^2 ||v||_{L^2}^2 + 2\lambda\mu ||u||_{L^2}||u||_{L^2} = (\lambda||u||_{L^2} + \mu||v||_{L^2})^2 \leq \\ \leq \lambda||u||_{L^2}^2 + \mu||v||_{L^2}^2 = \lambda \int_0^T u^2 dt + \mu \int_0^T v^2 dt,\end{aligned}$$

where the first inequality is the Cauchy-Schwarz inequality and the last inequality is due to convexity of the square function. To get equality when $0 < \lambda < 1$ in the first inequality we need $u = kv$, or for either u or v to be the zero function. To get equality in the last inequality we get $||u||_{L^2} = ||v||_{L^2}$ since x^2 is strictly convex. Thus for equality we need $u = \pm v$, but both u and $-u$ cannot belong to X unless $u \equiv 0$. Either way we get $u = v$. Hence we have strict convexity. This shows that $x \mapsto \int_0^T \ddot{x}^2 dt$ is convex, but not strictly since functions can differ by a linear function

and still have the same second derivative. Convexity of the sum part of f follows from convexity of the square function.

This shows that f is convex, now we want to show that it is strictly convex. We want to show that

$$f(\lambda u + (1-\lambda)v) = \lambda f(u) + (1-\lambda)f(v) \implies u = v.$$

Since both the integral and the sum is convex, we need equality both for the integral and the sum. Due to the strict convexity shown above, we need $\ddot{u} = \ddot{v}$. Thus $v = u + At + B$. Choosing two square terms from the sum, we get from strict convexity that at these points, say t_1 and t_2, we need $u(t_1) = u(t_2)$ and $v(t_1) = v(t_2)$. But then $A = B = 0$ and the strict convexity follows.

□

2.2 Functions of bounded variation, Stieltjes integrals and classification of the dual space of $C([a,b])$

We now introduce a class of functions called functions of bounded variation. These will together with a generalization of the Riemann integral be used to classify all bounded functionals on the space of continuous functions.

Definition 2.3. The space $BV[a,b]$ is called the space of bounded variation and consists of all functions on $[a,b] \to \mathbb{R}$ such that the total variation is finite. The total variation is defined as

$$T.V.(f) = \sup \sum_{i=1}^{n} |f(x_i) - f(x_{i-1})|$$

where the supremum is taken over all finite partitions $a = x_0 < x_1 < ... < x_{n-1} < x_n = b$ of the interval.

The norm of an element in $BV[a,b]$ is

$$||f||_{BV[a,b]} = |f(a)| + T.V.(f)$$

Essentially, the space consists of functions that do not change their function values too much. An important subclass is the normalized space of functions of bounded variation. This space removes some of the functions, that for our purposes are too similar.

Definition 2.4. The normalized space of functions of bounded variation denoted $NBV[a,b]$ consists of all functions of bounded variation on $[a,b]$ which vanish at the point a and which are continuous from the right in (a,b). The norm of an element in this space is $||v|| = T.V.(v)$.

We use a generalization of the Riemann integral called Riemann-Stieltjes integeral, or only Stiltjes integral. The theory behind this is in many ways similar to the Riemann integral. Apart from the integrand $f(x)$, we now have an additional function $\alpha(x)$ of bounded variation. Similar to the Riemann integral we partition the interval we integrate over into several subintervals and approximate $f(x)$ with some value it obtains on these subintervals. However, instead of multiplying this function value with the length of the subinterval, we multiply with the difference in α-values between the endpoints.

More formally, let $a = t_1 < t_2 < \cdots < t_n = b$ be a partition of $[a, b]$. Then we define the integral

$$\int_a^b f(x)d\alpha(x)$$

as the limit of

$$\sum_{i=1}^{n-1} f(c_i)(\alpha(t_{i+1}) - \alpha(t_i))$$

as the largest distance between the points in the partition goes to 0. The value c_i lies between t_i and t_{i+1}. This value exists when α is of bounded variation and f is continuous.

If the function α is differentiable we have

$$\int_a^b f d\alpha = \int_a^b f\alpha' dx. \tag{5}$$

We also have integration by parts

$$\int_a^b f d\alpha = f(b)\alpha(b) - f(a)\alpha(a) - \int_a^b \alpha df. \tag{6}$$

If α and f are differentiable this can be shown using (5) together with integration by parts as usual. However, this is true even when these additional restrictions on f and α do not hold. See Appell et al. [9].

The following theorem shows how we can characterize the dual space of $C([a, b])$.

Theorem 5. *Let f be a bounded linear functional on on $X = C[a, b]$. Then there is a function v of bounded variation on $[a, b]$ such that for all $x \in X$*

$$f(x) = \int_a^b x(t)dv(t)$$

and such that the norm of f is the total variation of v on $[a, b]$. Conversely, every function of bounded variation on $[a, b]$ defines a bounded linear functional on X in this way.

The representation of $C[a, b]$ becomes unique if we instead use functions in $NBV[a, b]$ [4].

2.3 The Karush-Kuhn-Tucker (KKT) conditions

We will soon state an infinite dimensional version of the KKT conditions, but first we need some more definitions. This theory can be found in Luenberger [4].

We will first define a way to talk about inequalities in vector spaces. To do that we first need to define a cone.

Definition 2.5. Let X be a vector space. A set $P \subset X$ is a cone if $x \in P$ implies that $\alpha x \in P$, for positive numbers α.

These can be used to define inequalities between elements in a vector space.

Definition 2.6. Let P be a convex cone in a normed vector space X. For $x, y \in X$, we write $x \geq y$ (with respect to P) if $x - y \in P$. We write $x > y$ if $x - y$ is in the interior of P. The cone P defining this relation is called the *positive cone* in X.

Throughout the report we will for $C([0,T])$ use the positive cone of functions that are non-negative on the entire interval.

When we have a positive cone in a normed space X we obtain a positive cone in the dual X^* defined as

$$P^* = \{x^* \in X^* : \langle x, x^* \rangle \geq 0 \quad \forall x \in P\}.$$

A generalization of directional derivatives to vector spaces are the Gateaux derivatives.

Definition 2.7. Let X be a vector space, Y a normed space and $T(x)$ a transformation defined on $D \subset X$ with range $R \subset Y$. Let $x \in D$ and let $h \in X$. If the limit

$$\delta T(x; h) = \lim_{\alpha \to 0} \frac{1}{\alpha}(T(x + \alpha h) - T(x)) \tag{7}$$

exists, it is called the Gateaux differential of T at x with increment h. If the limit (7) exists for all $h \in X$, the transformation T is said to be Gateaux differentiable at x.

When T is a functional, i.e. $Y = \mathbb{R}$, this can be written using ordinary derivatives as

$$\delta f(x, h) = \frac{d}{d\alpha} f(x + \alpha h)|_{\alpha=0}.$$

We can also define stationary points for functionals, similar to how these are defined for ordinary derivatives.

Definition 2.8. Let f be a functional defined on a vector space X, and suppose it is Gateaux differentiable. We say that f has a stationary point in x_0 if

$$\delta f(x_0, h) = 0,$$

for all $h \in X$.

Stationary points in vector spaces have some similar properties as stationary points in $\mathbb{R}^n$. For example, if a real valued function is defined on a vector space X and is Gateaux differentiable everywhere, then a necessary condition for a point to be a local extreme point is that it is a stationary point.

We need one more definition before we are ready to state the KKT-conditions for vector spaces.

Definition 2.9. Let X be a vector space and let Z be a normed space with a positive cone P having nonempty interior. Let G be a mapping $G : X \to Z$ which has a Gateaux differential that is linear in its increment. A point $x_0 \in X$ is said to be a *regular point* of the inequality $G(x) \leq 0$ if $G(x_0) \leq 0$ and there is an $h \in X$ such that $G(x_0) + \delta G(x_0; h) < 0$.

We now use a variant of the KKT-conditions applicable to vector spaces. The following theorem is taken from Luenberger [4].

Theorem 6 (Generalized Karush-Kuhn-Tucker theorem)**.** *Let X be a vector space and Z a normed space having a positive cone P. Assume that P contains an interior point.*
Let f be a Gateaux differentiable real-valued functional on X and G a Gateaux differentiable mapping from X into Z. Assume that the Gateaux differentials are linear in their increments. Suppose that x_0 minimizes f subject to $G(x) \leq 0$ and that x_0 is a regular point of the inequality $G(x) \leq 0$. Then there is a $z_0^* \in Z^*$, $z_0^* \geq 0$ such that the Lagrangian

$$f(x) + \langle G(x), z_0^* \rangle$$

is stationary at x_0; furthermore, $\langle G(x_0), z_0^* \rangle = 0$.

Comparing this to the finite dimensional KKT-conditions, see for example [6], we can see some differences and similarities. First of all this version only deals with inequality constraints. One could think that an equality constraint could be replaced with two inequality constraints, but then no point would be regular. The regularity also excludes choosing the positive cone as a point to force equality.

Otherwise the concept is rather similar. We add something to our function that forces a minimizer to become a stationary point of the Lagrangian. We now add $\langle G(x), z_0^* \rangle$, which may look a bit different from Lagrange multipliers. But in the case of a finite dimensional space, linear functionals are precisely dot products.

We also have the counterpart to the complementary slackness condition with $\langle G(x_0), z_0^* \rangle = 0$, and that $z_0^* \geq 0$. Once again this is exactly the same in finite dimensions.

In order to apply this theorem on our problem we need to choose the different sets and functions used, in order to fit our problem.

First of all we define the space X. We want to have the inequalities in G so we let $X = \{x \in H^2((0,T)), x(0) = 0\}$. We choose $Z = C([0,T]) \times \mathbb{R}$ with norm

$$||(w,\alpha)|| = (||w||_\infty^2 + |\alpha|^2)^{1/2}$$

Let $G(x) = (-\dot{x}, x(T) - x_{max})$. Since $\dot{x} \in H^1((0,T)) \subset C([0,T])$, we have $G(x) \in Z$.

We use the positive cone $P = \{(w,\alpha) \in Z; w \geq 0, \alpha \geq 0\}$. Then we see that $G(x) \leq 0$ is exactly the inequalities $\dot{x} \geq 0$ and $x(T) \leq x_{max}$.

One can also see that P has interior points, for example the function $g(x) = 1$ together with $\frac{x_{max}}{2}$.

According to Theorem 5 we can identify Z^* with $NBV([0,T]) \times \mathbb{R}$. We now give a description of the positive cone in Z^*.

Lemma 2. *The positive cone P^* in Z^* is*

$$P^* = \{(\nu,\mu) \in NBV([0,T]) \times \mathbb{R}; \nu \text{ is nondecreasing and } \mu \geq 0\}.$$

Proof. To show that this is the case, if $z \in P$ and $z^* \in P^*$ then $\langle z, z^* \rangle \geq 0$. On the other hand we can find elements in $z \in P$ such that if $z^* \in Z^* \setminus P^*$ then $\langle z, z^* \rangle < 0$. First of all, suppose we have $\mu < 0$, then we can choose the element z to be the zero function and the number 1. Then $\langle z, z^* \rangle = -\mu$.

On the other hand if ν, is not nondecreasing on the whole interval, then there exists $0 \leq a_1 < b_1 \leq T$ such that $\nu(b_1) - \nu(a_1) = -4\delta < 0$. We will assume $b_1 < T$, as the case $b_1 = T$ can be treated similarly. Due to right continuity, we can find a_2, b_2 with $a_1 < a_2 < b_1 < b_2 < T$, such that for any $a \in [a_1, a_2]$, $b \in [b_1, b_2]$ we have $\nu(b) - \nu(a) < -3\delta$. Since ν is of bounded variation the total variation, V, of ν on the interval $[a_1, a_2]$ is finite. Divide this interval into n subintervals. Since the total variation on each subinterval must sum to V we can conclude that at least one subinterval must have a total variation of at most $\frac{V}{n}$. Choosing n large enough we can find an interval $[A_1, A_2]$ with total variation less than δ. We define B_1, B_2 similarly. We define the function $g(t)$ that is zero outside of $[A_1, B_2]$, has the value 1 on $[A_2, B_1]$ and such that the graph of g is a straight line on $[A_1, A_2]$ and $[B_1, B_2]$, so that $g(t)$ is continuous.

Let $0 = t_0 < t_1 < \cdots < t_{k_1} = A_1 < \cdots < t_{k_2} = A_2 < \cdots < t_{k_3} = B_1 < \cdots < t_{k_4} = B_2 < \cdots < t_n = T$ be a partition of $[0,T]$, which includes the points A_1, A_2, B_1, B_2. Let $c_i \in (t_i, t_{i+1})$. Then $g(c_i) = 0$ outside of the interval $[A_1, B_2]$ and $g(c_i) \leq 1$ inside

the interval. We have

$$\sum_{i=0}^{n-1} g(c_i)\left(\nu(t_{i+1})-\nu(t_i)\right) = \sum_{i=0}^{k_1-1} g(c_i)\left(\nu(t_{i+1})-\nu(t_i)\right) + \sum_{i=k_1}^{k_2-1} g(c_i)\left(\nu(t_{i+1})-\nu(t_i)\right) +$$
$$+\sum_{i=k_2}^{k_3-1} g(c_i)\left(\nu(t_{i+1})-\nu(t_i)\right) + \sum_{i=k_3}^{k_4-1} g(c_i)\left(\nu(t_{i+1})-\nu(t_i)\right) + \sum_{i=k_4}^{n-1} g(c_i)\left(\nu(t_{i+1})-\nu(t_i)\right) \leq$$
$$\leq \sum_{i=k_1}^{k_2-1} g(c_i)|\nu(t_{i+1})-\nu(t_i)| + \sum_{i=k_2}^{k_3-1} 1\cdot\left(\nu(t_{i+1})-\nu(t_i)\right) + \sum_{i=k_3}^{k_4-1} g(c_i)|\nu(t_{i+1})-\nu(t_i)| \leq$$
$$\leq \sum_{i=k_1}^{k_2-1} |\nu(t_{i+1})-\nu(t_i)| + \nu(B_1) - \nu(A_2) + \sum_{i=k_3}^{k_4-1} |\nu(t_{i+1})-\nu(t_i)| \leq$$
$$\leq \delta - 3\delta + \frac{\delta}{3} = -\delta < 0.$$

This holds for any such partition and it follows that

$$\int_0^T g(t)d\nu(t) < 0.$$

□

The Gateaux differential of G is

$$\lim_{\alpha\to 0}\frac{G(x+\alpha h)-G(x)}{\alpha} =$$
$$= \lim_{\alpha\to 0}\frac{(-\dot{x}-\alpha\dot{h}, x(T)+\alpha h(T)-x_{max}) - (-\dot{x}, x(T)-x_{max})}{\alpha} =$$
$$= \lim_{\alpha\to 0}\frac{(-\alpha\dot{h}, \alpha h(T))}{\alpha} = (-\dot{h}, h(T)).$$

This is clearly linear in the increment.

We now compute the Gateuax differential of f.

$$\lim_{\alpha\to 0}\frac{1}{2\alpha}\left(\int_0^T(\ddot{x}+\alpha\ddot{h})^2dt - \int_0^T\ddot{x}^2dt + \sum_{i=1}^m(x(t_i)+\alpha h(t_i)-\alpha_i)^2 - \sum_{i=1}^m(x(t_i)-\alpha_i)^2\right) =$$
$$= \lim_{\alpha\to 0}\frac{1}{2}\left(\int_0^T \alpha\ddot{h}^2 + 2\ddot{x}\ddot{h}dt + \sum_{i=1}^m h(t_i)(\alpha h(t_i)+2x(t_i)-2\alpha_i)\right) =$$
$$= \int_0^T \ddot{x}\ddot{h}dt + \sum_{i=1}^m h(t_i)(x(t_i)-\alpha_i)$$

This is also linear in its increments.

In order to use the KKT-conditions we need to decide which points are regular. The following lemma shows that all points are regular.

Lemma 3. *Every $x \in X$ with $G(x) \leq 0$ is a regular point of the inequality $G(x) \leq 0$.*

Proof. We need to show that if $G(x) \leq 0$ there is an $h \in X$ such that

$$(-\dot{x}, x(T) - x_{max}) + (-\dot{h}, h(T)) = (-\dot{x} - \dot{h}, x(T) + h(T) - x_{max}) < 0.$$

This is equivalent to

$$\begin{cases} -\dot{x} - \dot{h} < 0 \\ x(T) + h(T) - x_{max} < 0. \end{cases}$$

This can be achieved by letting $h = \frac{x_{max}}{2T}t - x(t)$. Then $h \in X$ and both inequalities are true, which finishes the proof. □

Lemma 4. *Let $x_* \in X$ be the minimizer of the minimization problem* (1)*, and denote $u_* = \ddot{x}_*$. Then u_* is affine on each subinterval of $[t_{i-1}, t_i)$ where $\dot{x}_* > 0$.*

Proof. By using the KKT-conditions, there exists $s^* = (\nu_*, \mu_*) \in P^*$, i.e. $\nu_* \in NBV([0,T])$ and $\mu_* \in \mathbb{R}$, with $\mu_* \geq 0$ and ν_* non-decreasing, such that

$$\int_0^T \ddot{x}_*(t)\ddot{h}(t)dt + \sum_{i=1}^{m}(x_*(t_i) - \alpha_i)h(t_i) - \int_0^T \dot{h}(t)d\nu_*(t) + \mu_* h(T) = 0, \tag{8}$$

for all $h \in X$. We also get

$$-\int_0^T \dot{x}_*(t)d\nu_* + \mu_*(x_*(T) - x_{max}) = 0. \tag{9}$$

Due to the conditions on x_*, ν_* and μ_* we get

$$-\int_0^T \dot{x}_*(t)d\nu_* \leq 0 \quad \text{and} \quad \mu_*(x_*(T) - x_{max}) \leq 0,$$

which together with equation (9) shows that both terms has to equal zero. For the integral this means that if $\dot{x}_* > 0$ on an interval, then ν_* has to be constant on this interval. Otherwise this part would contribute with a positive value, and no other part of the integral can contribute with a negative value.

The second integral in equation (8) can be rewritten, using integration by parts, as

$$\int_0^T \dot{h}(t)d\nu_*(t) = \dot{h}(T)\nu_*(T) - \dot{h}(0)\nu_*(0) - \int_0^T \nu_*(t)d\dot{h}(t)$$

Since $\dot{h}$ is differentiable the last integral can be rewritten as $\int_0^T \nu_*(t)\ddot{h}(t)dt$. Thus we can rewrite equation (8) and get

$$\int_0^T (\ddot{x}_*(t)+\nu_*)\ddot{h}(t)dt+\sum_{i=1}^{m} w_i(x_*(t_i)-\alpha_i)h(t_i)+\dot{h}(0)\nu_*(0)-\dot{h}(T)\nu_*(T)+\mu_* h(T)=0.$$

In particular this holds for h that is zero everywhere except in the interior of one of the intervals. This gives us

$$\int_{t_i}^{t_{i+1}} (\ddot{x}_*(t)+\nu_*)\ddot{h}(t)dt=0$$

for all $h \in C_0^\infty(t_i, t_{i+1})$. Then it can be shown that $\ddot{x}_*(t)+\nu_*$ is affine on this interval, see e.g. Hörmander [11]. If $\dot{x}_* > 0$ on this interval then ν_* is constant here and hence $\ddot{x}_*$ is affine on this interval.

□

2.4 Derivation of the optimal function on each subinterval

These results can now be used to find the optimal curve on an interval, given that the values and the derivatives are known at both endpoints. When a third degree polynomial is increasing this will still be optimal also when the monotinicity constraint is added, but when it is not, we need another type of curve. We first find a necessary and sufficient condition for when the third degree polynomial is increasing.

Theorem 7. *Let $P(t)$ be a third degree polynomial defined on the interval $[0, t_F]$, with $P(0)=0$, $P(t_F)=x_F$, $P'(0)=\dot{x}_0$ and $P'(t_F)=\dot{x}_F$. This polynomial is increasing if and only if*

$$x_F \geq \frac{t_F}{3}\left(\dot{x}_0+\dot{x}_F-\sqrt{\dot{x}_0\dot{x}_F}\right).$$

Proof. The third degree polynomial that goes through the points (0, 0), (t_F, x_F) and has a derivative that passes through $(0, \dot{x}_0)$ and $(t_F, \dot{x}_F)$ is

$$P(t)=\left(\frac{\dot{x}_0+\dot{x}_F}{t_F^2}-\frac{2x_F}{t_F^3}\right)t^3+\left(\frac{3x_F}{t_F^2}-\frac{2\dot{x}_0+\dot{x}_F}{t_F}\right)t^2+\dot{x}_0 t,$$

with derivative

$$P'(t)=3\left(\frac{\dot{x}_0+\dot{x}_F}{t_F^2}-\frac{2x_F}{t_F^3}\right)t^2+2\left(\frac{3x_F}{t_F^2}-\frac{2\dot{x}_0+\dot{x}_F}{t_F}\right)t+\dot{x}_0.$$

We want to find exactly when $P'(t) \geq 0$ for all $t \in [0, t_F]$. To simplify the calculations, we will use $P'(t)=At^2+Bt+C$. If $A \leq 0$ then $P'(t) \geq 0$ on the interval since both $P'(0)$ and $P'(t_F)$ are non-negative. Thus P is increasing if $x_F \geq \frac{t_F}{2}(\dot{x}_0+\dot{x}_F)$.

Now suppose that $A > 0$, i.e. $x_F < \frac{t_F}{2}(\dot{x}_0 + \dot{x}_F)$. The smallest value of $P'(t)$ is obtained at $t = -\frac{B}{2A}$. If this point lies outside the interval then the function is positive since it is positive at the edges of the interval. This happens if

$$-\frac{B}{2A} \leq 0 \iff B \geq 0 \iff x_F \geq \frac{t_F}{3}(2\dot{x}_0 + \dot{x}_F)$$

or if

$$-\frac{B}{2A} \geq t_F \iff B + 2At_F \leq 0 \iff x_F \geq \frac{t_F}{3}(\dot{x}_0 + 2\dot{x}_F).$$

If $0 < -\frac{B}{2A} < t_F$, i.e. $x_F < \frac{t_F}{3}\min\{\dot{x}_0 + 2\dot{x}_F, 2\dot{x}_0 + \dot{x}_F\}$, then the function is positive if and only if $P'(-\frac{B}{2A}) \geq 0$.

$$P'(-\frac{B}{2A}) = C - \frac{B^2}{4A} \geq 0 \iff 4AC - B^2 = \frac{4\dot{x}_0\dot{x}_F}{t_F^2} - 4\left(\frac{3x_F}{t_F^2} - \frac{\dot{x}_0 + \dot{x}_F}{t_F}\right)^2 \geq 0 \iff$$
$$\iff \frac{t_F}{3}(\dot{x}_0 + \dot{x}_F - \sqrt{\dot{x}_0\dot{x}_F}) \leq x_F \leq \frac{t_F}{3}(\dot{x}_0 + \dot{x}_F + \sqrt{\dot{x}_0\dot{x}_F})$$

The upper inequality is fulfilled since $x_F < \frac{t_F}{3}\min\{\dot{x}_0 + 2\dot{x}_F, 2\dot{x}_0 + \dot{x}_F\} \leq \frac{t_F}{3}(\dot{x}_0 + \dot{x}_F + \sqrt{\dot{x}_0\dot{x}_F})$. Therefore, in this case, we get $x_F \geq \frac{t_F}{3}(\dot{x}_0 + \dot{x}_F - \sqrt{\dot{x}_0\dot{x}_F})$ for the function to be positive.

If $x_F \geq \frac{t_F}{3}(\dot{x}_0 + \dot{x}_F - \sqrt{\dot{x}_0\dot{x}_F})$ then one of the cases above shows that the function is increasing. On the other hand if the function is increasing, one of the inequalities has to hold and the smallest value x_F can take is $\frac{t_F}{3}(\dot{x}_0 + \dot{x}_F - \sqrt{\dot{x}_0\dot{x}_F})$.

□

Theorem 8. *The minimizer of*

$$\int_0^{t_F} \ddot{x}(t)^2 dt$$

subject to

$$\begin{cases} x \in H^2(0, t_F) \\ \dot{x}(t) \geq 0 \text{ for } 0 \leq t \leq t_F \end{cases} \quad \text{and} \quad \begin{cases} x(0) = 0 \\ x(t_F) = x_F \\ \dot{x}(0) = \dot{x}_0 \\ \dot{x}(t_F) = \dot{x}_F \end{cases}$$

is

$$f_1(t) = \dot{x}_0 t + \left(\frac{3x_F}{t_F^2} - \frac{2\dot{x}_0 + \dot{x}_F}{t_F}\right)t^2 + \left(\frac{\dot{x}_0 + \dot{x}_F}{t_F^2} - \frac{2x_F}{t_F^3}\right)t^3$$

if $x_F \geq \frac{t_F}{3}(\dot{x}_0 + \dot{x}_F - \sqrt{\dot{x}_0\dot{x}_F})$, *and otherwise*

$$f_2(t) = \begin{cases} \frac{x_F \dot{x}_0^{3/2}}{\dot{x}_0^{3/2}+\dot{x}_F^{3/2}} + \frac{(\dot{x}_0^{3/2}+\dot{x}_F^{3/2})^2}{27x_F^2}\left(t - \frac{3x_F\dot{x}_0^{1/2}}{\dot{x}_0^{3/2}+\dot{x}_F^{3/2}}\right)^3 & \text{if} \quad 0 \le t < \frac{3x_F\dot{x}_0^{1/2}}{\dot{x}_0^{3/2}+\dot{x}_F^{3/2}} \\ \frac{3x_F\dot{x}_0^{3/2}}{\dot{x}_0^{3/2}+\dot{x}_F^{3/2}} & \text{if} \quad \frac{3x_F\dot{x}_0^{1/2}}{\dot{x}_0^{3/2}+\dot{x}_F^{3/2}} \le t \le t_F - \frac{3x_F\dot{x}_F^{1/2}}{\dot{x}_0^{3/2}+\dot{x}_F^{3/2}} \\ \frac{x_F \dot{x}_0^{3/2}}{\dot{x}_0^{3/2}+\dot{x}_F^{3/2}} + \frac{(\dot{x}_0^{3/2}+\dot{x}_F^{3/2})^2}{27x_F^2}\left(t - t_F + \frac{3x_F\dot{x}_F^{1/2}}{\dot{x}_0^{3/2}+\dot{x}_F^{3/2}}\right)^3 & \text{if} \quad t_F - \frac{3x_F\dot{x}_F^{1/2}}{\dot{x}_0^{3/2}+\dot{x}_F^{3/2}} < t \le t_F. \end{cases}$$

The value of the integral is

$$\int_0^{t_F} (f_i'')^2 = \begin{cases} 4\frac{(\dot{x}_0^2+\dot{x}_F^2)(\Delta t)^2 - 3\Delta x(\dot{x}_0+\dot{x}_F)\Delta t + 3(\Delta x)^2 + \dot{x}_0\dot{x}_F(\Delta t)^2}{(\Delta t)^3} & \text{if } i = 1 \\ \frac{4}{9\Delta x}\left(\dot{x}_0^{3/2} + \dot{x}_F^{3/2}\right)^2 & \text{if } i = 2. \end{cases}$$

If the curve starts in another point (t_0, x_0) then the curve is moved correctly by adding x_0, replacing t with $t - t_0$, t_F by $t_F - t_0$ and x_F with $x_F - x_0$.

Worth noting is that the second type of function is only a function if $\Delta x \le \frac{\Delta t}{3}(\dot{x}_0 + \dot{x}_F - \sqrt{\dot{x}_0\dot{x}_F})$, otherwise the intervals will overlap. The function is continuously differentiable.

Proof. We have already shown that the third degree polynomial is increasing if $x_F \ge \frac{t_F}{3}(\dot{x}_0 + \dot{x}_F - \sqrt{\dot{x}_0\dot{x}_F})$. This will be optimal as shown in Theorem 1.

We will now continue to show that when $x_F < \frac{t_F}{3}(\dot{x}_0 + \dot{x}_F - \sqrt{\dot{x}_0\dot{x}_F})$ we get the second type of function. By Lemma 4 we know that if there are no points for which $\dot{x}_*$ is 0 then the $\ddot{x}_*$ is affine and the optimal function is a third degree polynomial. But as could be seen above this does not give us an increasing function in this case. Thus there has to be some point or interval where $\dot{x}_* = 0$. Let us assume that this interval is $[t_1, t_2]$ and that $x_* = A$ on this interval. In order to get an admissible function we need that $\dot{x}(t_1) = \dot{x}(t_2) = 0$. We can now try to fit the optimal curve satisfying these conditions, i.e, one third degree polynomial on $[0, t_1]$ and another on $[t_2, t_F]$. For this to become a function we need $0 < t_1 \le t_2 < t_F$. We will however, not include the boundary $t_1 \le t_2$ for now since, as will be apparent later, this is not necessary. One can also note that we also include functions where the interval on which $\dot{x}_* = 0$ is not really an interval but rather just a point, which happens when $t_1 = t_2$.

The value of the integral we want to minimize becomes, as a function of the parameters t_1, t_2, A,

$$f(A, t_1, t_2) = \frac{4}{t_1^3}(\dot{x}_0^2 t_1^2 - 3A\dot{x}_0 t_1 + 3A^2) + \\ + \frac{4}{(t_f - t_2)^3}(\dot{x}_0^2(t_f - t_2)^2 - 3(x_F - A)\dot{x}_0(t_F - t_2) + 3(x_F - A)^2).$$

The gradient of this function is

$$\nabla f = \begin{bmatrix} \frac{4}{t_1^3}(6A - 3\dot{x}_0 t_1) + \frac{4}{(t_F - t_2)^3}(3\dot{x}_f(t_F - t_2) - 6(x_F - A)) \\ -\frac{4\dot{x}_0^2}{t_1^2} + \frac{24A\dot{x}_0}{t_1^3} - \frac{36A^2}{t_1^4} \\ \frac{4\dot{x}_F^2}{(t_F - t_2)^2} - \frac{24(x_F - A)\dot{x}_F}{(t_F - t_2)^3} + \frac{36(x_f - A)^2}{(t_F - t_2)^4} \end{bmatrix}.$$

We now find stationary points by finding points where the gradient is zero. This gives us the following system of equations

$$\begin{cases} \frac{4}{t_1^3}(6A - 3\dot{x}_0 t_1) + \frac{4}{(t_F - t_2)^3}(3\dot{x}_f(t_F - t_2) - 6(x_F - A)) & = 0 \\ -\frac{4\dot{x}_0^2}{t_1^2} + \frac{24A\dot{x}_0}{t_1^3} - \frac{36A^2}{t_1^4} & = 0 \\ \frac{4\dot{x}_F^2}{(t_F - t_2)^2} - \frac{24(x_F - A)\dot{x}_F}{(t_F - t_2)^3} + \frac{36(x_f - A)^2}{(t_F - t_2)^4} & = 0 \end{cases} \tag{10}$$

The second and third equation can be rewritten as a square, giving us

$$\begin{cases} (\frac{\dot{x}_0}{t_1} - \frac{3A}{t_1^2})^2 & = 0 \\ (\frac{\dot{x}_F}{t_F - t_2} - \frac{3(x_F - A)}{(t_F - t_2)^2})^2 & = 0 \end{cases}$$

showing that

$$\begin{cases} \dot{x}_0 t_1 & = 3A \\ \dot{x}_F(t_F - t_2) & = 3(x_F - A). \end{cases}$$

With this the first equation in (10) can be rewritten as

$$\dot{x}_0(t_F - t_2)^2 = \dot{x}_F t_1^2,$$

and with some substitutions we get

$$(\dot{x}_F^3 - \dot{x}_0^3)t_1^2 + 6x_F\dot{x}_0^2 t_1 - 9x_F^2\dot{x}_0 = 0.$$

Assuming $\dot{x}_F \neq \dot{x}_0$ we can solve for t_1 and get

$$t_1 = -\frac{3x_F\dot{x}_0^2}{\dot{x}_F^3 - \dot{x}_0^3} \pm \sqrt{\left(\frac{3x_F\dot{x}_0^2}{\dot{x}_F^3 - \dot{x}_0^3}\right)^2 + \frac{9x_F^2\dot{x}_0}{\dot{x}_F^3 - \dot{x}_0^3}}.$$

Simplifying we get

$$t_1 = \frac{3x_F\dot{x}_0^{1/2}}{\dot{x}_F^{3/2} + \dot{x}_0^{3/2}} \quad \text{or} \quad t_1 = \frac{3x_F\dot{x}_0^{1/2}}{\dot{x}_0^{3/2} - \dot{x}_F^{3/2}}.$$

The second solution for t_1 makes either $t_1 < 0$ or $t_2 > t_F$ leaving us with only the first solution. Then

$$\begin{cases} t_2 &= t_F - \frac{3x_F\dot{x}_0^{1/2}}{\dot{x}_0^{3/2}+\dot{x}_F^{3/2}} \\ A &= \frac{x_F\dot{x}_0^{3/2}}{\dot{x}_F^{3/2}+\dot{x}_0^{3/2}} \end{cases}$$

If $\dot{x}_F = \dot{x}_0$ we get the same expressions.

For both the intervals $[0, t_1]$ and $[t_2, t_F]$ we have that $\Delta x \geq \frac{\Delta t}{3}(v_l + v_r + \sqrt{v_l v_r})$, and the third degree polynomials are thus increasing.

We can also see that $t_1 \leq t_2$, or

$$\frac{3x_F\dot{x}_0}{\dot{x}_0^{3/2} + \dot{x}_F^{3/2}} \leq t_F - \frac{3x_F\dot{x}_F}{\dot{x}_0^{3/2} + \dot{x}_F^{3/2}}$$

whenever $\Delta x < \frac{\Delta t}{3}(v_l + v_r + \sqrt{v_l v_r})$. This holds since if $x_F \leq \frac{t_F}{3}(\dot{x}_0 + \dot{x}_F - \sqrt{\dot{x}_0\dot{x}_F})$ then

$$\frac{3x_F\dot{x}_0}{\dot{x}_0^{3/2} + \dot{x}_F^{3/2}} + \frac{3x_F\dot{x}_F}{\dot{x}_0^{3/2} + \dot{x}_F^{3/2}} \leq \frac{3\frac{t_F}{3}(\dot{x}_0 + \dot{x}_F - \sqrt{\dot{x}_0\dot{x}_F})\dot{x}_0}{\dot{x}_0^{3/2} + \dot{x}_F^{3/2}} + \frac{3\frac{t_F}{3}(\dot{x}_0 + \dot{x}_F - \sqrt{\dot{x}_0\dot{x}_F})\dot{x}_F}{\dot{x}_0^{3/2} + \dot{x}_F^{3/2}} =$$

$$= \frac{t_F(\dot{x}_0 + \dot{x}_F - \sqrt{\dot{x}_0\dot{x}_F})(\dot{x}_0^{1/2} + \dot{x}_F^{1/2})}{\dot{x}_0^{3/2} + \dot{x}_F^{3/2}} = t_F$$

We see that the function value approaches ∞ as $t_1 \to 0$ or $t_2 \to t_f$. The same is true when A goes to $\pm\infty$ since the coefficient in front of A^2 is positive. Since the function is differentiable everywhere on our area, the found point must be the minimum. These values for t_1, t_2 and A gives us the function found in the theorem. □

3 Algorithm

In the previous section we found that given the values of the function and its derivative at the spline knots, we can find the optimal curve. Hence to find the optimal curve when these are not known, we can consider these values to be variables of an optimization problem. With

$$V(\Delta x, v_l, v_r, \Delta t) =$$

$$\begin{cases} 4\frac{(v_l^2+v_r^2)(\Delta t)^2-3\Delta x(v_l+v_r)\Delta t+3(\Delta x)^2+v_l v_r(\Delta t)^2}{(\Delta t)^3} & \text{if } \Delta x \geq \frac{\Delta t}{3}(v_l + v_r\sqrt{v_l v_r}) \\ \frac{4}{9\Delta x}\left(v_l^{3/2} + v_r^{3/2}\right)^2 & \text{otherwise,} \end{cases} \tag{11}$$

we thus want to minimize

$$\sum_{i=1}^{n-1} V(x_{i+1} - x_i, v_i, v_{i+1}, t_{i+1} - t_i) + \frac{\lambda}{2}\sum_{i=1}^{n}(x_i - \alpha_i)^2, \tag{12}$$

subject to

$$\begin{cases} x_{i+1} - x_i \geq 0 \\ v_i \geq 0 \\ x_0 = 0 \\ x_n \leq x_{max} \end{cases}$$

The variables x_i and v_i corresponds to the value of the function and the derivative, respectively, at spline knot i. All t_i and α_i are fixed.

The function (11) is a piecewise defined function and will cause the function (12) to be defined differently in different regions of the $\mathbb{R}^{2n}$-space. The regions are separated by surfaces for which $x_{i+1} - x_i = \frac{t_{i+1}-t_i}{3}(v_{i+1} + v_i - \sqrt{v_{i+1}v_i})$. We refer to these as interfaces between regions.

However minimizing this seemed to give unstable results when using a numerical solver for optimization with constraints such as MATLABS's FMINCON on problems with more than 10 subintervals, and therefore an alternative algorithm was suggested in Maad Sasane [1]. This algorithm will be explored in this section. There will also be a section discussing the use of only FMINCON as it seems that one can improve the performance of this significantly.

The idea of the algorithm is to split the problem into subproblems, each of which are easier, and where one of them will have the correct solution. The subproblems are constructed by choosing to use only one of the rows in (11) per interval and

disregarding the additional conditions of the variables. For each of these subproblems we find the minimizer and checks if this solution also is a solution to the original problem. This is done by checking for each subinterval that $\Delta x \geq \frac{\Delta t}{3}(v_l + v_r - \sqrt{v_l v_r})$ if the interval is of type one and that we have the opposite inequality otherwise.

We will say that an interval is of type 1 if it uses the first line, and otherwise that it is of type 2. We start with the subproblem where all intervals are of type 1, then all subproblems where one interval is of type 2, then with two intervals of type 2 and so on. With n intervals, this gives us 2^n subproblems.

The following sections will prove that the algorithm is correct, as well as some additional properties, that can be used.

3.1 Convexity of the problems

We now show that our minimization problems are convex. This is highly useful since then we can be certain that we have a global minimum whenever we find a local minimum. To show convexity we will use Sylvester's criterion, see for example Böiers [6].

Lemma 5 (Sylvester's criterion)**.** *A symmetric matrix H is positive definite if and only if each of the leading principal minors are positive. If the determinant of the entire matrix is 0 instead of positive then the matrix is positive semidefinite.*

Theorem 9. *Each of the subproblems is strictly convex.*

Each of these functions is convex. Since the points t_i are fixed we only consider them as functions of the other variables.

Proof. The function in each subproblem consists of a sum of functions of the two types of functions defined in (11). We will show that each of these is convex by showing that the Hessian is positive semidefinite on the interior of the domain, i.e. when each of the variables are positive. Then by continuity we get that it is convex on the entire domain.

$$f(x, v_l, v_r) = 4\frac{(v_l^2 + v_r^2)(\Delta t)^2 - 3x(v_l + v_r)\Delta t + 3(x)^2 + v_l v_r(\Delta t)^2}{(\Delta t)^3}$$

We get the following gradient

$$\nabla f = \frac{4}{(\Delta t)^3}(-3\Delta t(v_l + v_r) + 6x, (\Delta t)^2(2v_l + v_r) - 3\Delta t x, (\Delta t)^2(v_l + 2v_r) - 3\Delta t x),$$

and the Hessian is

$$\frac{4}{(\Delta t)^3}\begin{bmatrix} 6 & -3\Delta t & -3\Delta t \\ -3\Delta t & 2(\Delta t)^2 & (\Delta t)^2 \\ -3\Delta t & (\Delta t)^2 & 2(\Delta t)^2 \end{bmatrix}.$$

We compute the leading principal minors (ignoring the factor $\frac{4}{(\Delta t)^3}$):

$$\left|\left[6\right]\right| = 6 > 0$$

and

$$\left|\begin{bmatrix} 6 & -3\Delta t \\ -3\Delta t & 2(\Delta t)^2 \end{bmatrix}\right| = 12(\Delta t)^2 - 9(\Delta t)^2 = 3(\Delta t)^2 > 0$$

since $\Delta t > 0$. The determinant of the entire matrix is

$$\left|\begin{bmatrix} 6 & -3\Delta t & -3\Delta t \\ -3\Delta t & 2(\Delta t)^2 & (\Delta t)^2 \\ -3\Delta t & (\Delta t)^2 & 2(\Delta t)^2 \end{bmatrix}\right| = 0.$$

The first two are strictly positive and the last is 0 which according to Lemma 5 shows that the Hessian is positively semidefinite on the interior and therefore the function is convex.

As for the second function,

$$g(x, v_l, v_r) = \frac{1}{x}\left(v_l^{3/2} + v_r^{3/2}\right)^2,$$

we get the gradient

$$\nabla g = \left(-\frac{1}{x^2}\left(v_l^{3/2} + v_r^{3/2}\right)^2, \frac{3\sqrt{v_l}}{x}\left(v_l^{3/2} + v_r^{3/2}\right), \frac{3\sqrt{v_r}}{x}\left(v_l^{3/2} + v_r^{3/2}\right)\right)$$

and Hessian

$$H = \begin{bmatrix} \frac{2}{x^3}\left(v_l^{3/2} + v_r^{3/2}\right)^2 & -\frac{3}{x^2}\sqrt{v_l}\left(v_l^{3/2} + v_r^{3/2}\right) & -\frac{3}{x^2}\sqrt{v_r}\left(v_l^{3/2} + v_r^{3/2}\right) \\ -\frac{3}{x^2}\sqrt{v_l}\left(v_l^{3/2} + v_r^{3/2}\right) & \frac{3}{x}(2v_l + \frac{1}{2}v_l^{-1/2}v_r^{3/2}) & \frac{9}{2x}\sqrt{v_l v_r} \\ -\frac{3}{x^2}\sqrt{v_r}\left(v_l^{3/2} + v_r^{3/2}\right) & \frac{9}{2x}\sqrt{v_l v_r} & \frac{3}{x}(2v_r + \frac{1}{2}v_l^{3/2}v_r^{-1/2}) \end{bmatrix}.$$

Once again we compute the leading principal minors

$$\left|\left[\frac{2}{x^3}\left(v_l^{3/2} + v_r^{3/2}\right)^2\right]\right| > 0,$$

on the interior.

$$\left|\begin{bmatrix} \frac{2}{x^3}\left(v_l^{3/2}+v_r^{3/2}\right)^2 & -\frac{3}{x^2}\sqrt{v_l}\left(v_l^{3/2}+v_r^{3/2}\right) \\ -\frac{3}{x^2}\sqrt{v_l}\left(v_l^{3/2}+v_r^{3/2}\right) & \frac{3}{x}(2v_l+\frac{1}{2}v_l^{-1/2}v_r^{3/2}) \end{bmatrix}\right| =$$

$$= \frac{6}{x^4}\left(v_l^{3/2}+v_r^{3/2}\right)^2\left(2v_l+\frac{1}{2}v_l^{-1/2}v_r^{3/2}\right) - \frac{9}{x^4}v_l\left(v_l^{3/2}+v_r^{3/2}\right)^2 =$$

$$= \frac{3}{x^4}v_l\left(v_l^{3/2}+v_r^{3/2}\right)^2 + \frac{6}{x^4}\left(v_l^{3/2}+v_r^{3/2}\right)^2\frac{1}{2}v_l^{-1/2}v_r^{3/2} > 0,$$

$$|H| = \frac{9}{2x^5}\left(v_l^{3/2}+v_r^{3/2}\right)^2\left(v_lv_r + v_l^{-1/2}v_r^{5/2} + v_l^{5/2}v_r^{-1/2}\right) > 0$$

Thus this is function is also convex on the interior of the domain.

In both of these functions we use $x_{i+1} - x_i$ instead of x but such a composition of an affine function and a convex function is still convex.

Each subproblem is a sum of the two types of functions above, using the appropriate variables for each interval and, as well as the sum $\sum_i (x_i - \alpha_i)^2$. Since the sum of convex functions are convex we get that each subproblem is convex.

We will now show that the subproblem is strictly convex. If f were a function of only v_l and v_r it would be strictly convex. Since the $(x_i - \alpha_i)^2$ is strictly convex and includes the x_i we get that the sum is strictly convex. □

Lemma 6. *The two types of functions have the same gradient when* $\Delta x = \frac{\Delta t}{3}(v_l + v_r - \sqrt{v_l v_r})$.

Proof. Consider the function

$$f(\Delta x, v_l, v_r) = \\ 4\frac{(v_l^2+v_r^2)(\Delta t)^2 - 3\Delta x(v_l+v_r)\Delta t + 3(\Delta x)^2 + v_lv_r(\Delta t)^2}{(\Delta t)^3} - \frac{4}{9\Delta x}\left(v_l^{3/2}+v_r^{3/2}\right)^2$$

that is the difference between the two types of cost functions. We will show that the gradient is zero when $\Delta x = \frac{\Delta t}{3}(v_l + v_r - \sqrt{v_l v_r})$.

$$f'_{\Delta x} = \frac{4}{(\Delta t)^3}(6\Delta x - 3(v_l+v_r)\Delta t) + \frac{4}{9(\Delta x)^2}(v_l^{3/2}+v_r^{3/2})^2.$$

On the interesting surface we get

$$f'_{\Delta x} = \frac{4}{(\Delta t)^3}(6\frac{\Delta t}{3}(v_l + v_r - \sqrt{v_l v_r}) - 3(v_l + v_r)\Delta t) +$$
$$+ \frac{4}{9(\frac{\Delta t}{3}(v_l + v_r - \sqrt{v_l v_r}))^2}(v_l^{3/2} + v_r^{3/2})^2 =$$
$$= \frac{4}{(\Delta t)^2}(\frac{(v_l^{3/2} + v_r^{3/2})^2}{(v_l + v_r - \sqrt{v_l v_r})^2} - (v_l + v_r + 2\sqrt{v_l v_r}) =$$
$$= \frac{4}{(\Delta t)^2(v_l + v_r - \sqrt{v_l v_r})^2}((v_l^{3/2} + v_r^{3/2})^2 - (v_l + v_r + 2\sqrt{v_l v_r})(v_l + v_r - \sqrt{v_l v_r})^2) = 0$$

Expanding the denominator we now get that this equals 0.

Similar calculations shows that $f'_{v_l} = 0$ at the same points:

$$f'_{v_l} = \frac{4}{(\Delta t)^3}(2v_l(\Delta t)^2 - 3\Delta x \Delta t + v_r(\Delta t)^2) - \frac{4}{9\Delta x}(3v_l^2 + 3v_l^{1/2}v_r^{3/2})$$

and replacing Δx we get

$$f'_{v_l} = \frac{4}{\Delta t}(2v_l - (v_l + v_r - \sqrt{v_l v_r}) + v_r) - \frac{4}{\Delta t(v_l + v_r - \sqrt{v_l v_r})}(v_l^2 + v_l^{1/2}v_r^{3/2}) =$$
$$= \frac{4}{\Delta t(v_l + v_r - \sqrt{v_l v_r})}((v_l + \sqrt{v_l v_r})(v_l + v_r - \sqrt{v_l v_r}) - v_l^2 - v_l^{1/2}v_r^{3/2}) = 0$$

By symmetry the same holds for v_r. Thus the two types of functions have the same derivative along these borders. □

This means that the function is differentiable everywhere.

Theorem 10. *The original problem* (12) *is convex.*

Proof. Clearly it is defined on a convex set. We have already shown that each of the two types of functions is convex. We also know that the gradients of the two types of functions are the same on the interface between the regions where they are defined. Thus if we restrict ourselves to a line then this function will have an increasing derivative on each side of the interface. But since the derivative is the same there, it is increasing everywhere. This holds for any restriction of the function to a line. Thus the problem is convex. □

3.2 Correctness of algorithm

This section will show that one of the subproblems will have the same solution as the original problem, and that whenever a solution to a subproblem is acceptable it will be the solution to the original problem as well.

We know that there exists a unique minimizer, x_0, to the original minimization problem. This solution will by necessity have each of its intervals belong to one of the two types of functions, corresponding to one of the subproblems. This subproblem and the original problem will be defined in exactly the same way, in a region containing x_0. In particular the function values and derivatives will be the same at x_0 causing this point to be a minima also for the subproblem. Due to strict convexity this will be the only minimum of the subproblem.

On the other hand, if a point which is a minimizer to a subproblem is acceptable, it will also be a minimizer of the original problem with the same argument. Due to strict convexity it will be the only solution here as well.

Since we visit all subproblems until we find an acceptable minimum point we can be sure that the solution will be found.

3.3 Bounding

This section will show that one can give some lower bounds for the minimum value of some subproblems given the solution to others.

Definition 3.1. A solution to a subproblem is called *acceptable* if it is feasible and satisfies the inequalities in (11) on each segment corresponding to the choice of function on that interval. That is for each interval $\Delta x \geq \frac{\Delta t}{3}(v_l + v_r - \sqrt{v_l v_r})$ if the interval is of type 1 and otherwise $\Delta x < \frac{\Delta t}{3}(v_l + v_r - \sqrt{v_l v_r})$.

We consider the set of subproblems as nodes in a graph consisting of several layers, where layer k contains all subproblems with k segments that are of type 2. Two subproblems in adjacent layers have a connection if they have the same type of functions on all intervals except one. We consider the highest layer to be the one containing no intervals of type two. This image motivates the following definition.

Definition 3.2. We say that a subproblem A is *below* another subproblem B if all intervals that are of type 2 in B also are of type 2 in A.

It should be noted that a subproblem A can be below both subproblems B and C, but neither B nor C is below the other.

Lemma 7. *On an interval, if* $\Delta x < \frac{\Delta t}{3}(v_l + v_r - \sqrt{v_l v_r})$ *then the function value of type 1 is smaller than the function value of type 2.*

Proof. When $\Delta x < \frac{\Delta t}{3}(v_l + v_r - \sqrt{v_l v_r})$ these variables correspond to a function. Since we know that the third degree polynomials will be optimal, all other functions will be worse. In particular the value of the type 2 function will be larger than the value of the type 1 function. □

Lemma 8. *Any acceptable solution below another solution has a larger value.*

Proof. Suppose branch b is below branch a, and b is acceptable. Let V_b be the value of b and V_a the value of a. If we keep the variables from the optimal solution in b but change the function from type 2 to type 1 on any interval where b is of type 2 but a is of type 1. Let V be the value obtained from this. Since b is acceptable, the values on these interval will be smaller in V than in V_b and all other intervals are the same. Thus $V_b < V$. We also know that V is obtainable in a and since V_a is optimal $V < V_a$. Hence $V_b < V_a$. □

Even though it is possible to use this to exclude some subproblems from the search, it requires us to already know a good value for the original problem. This is not however something we can expect to have. To get this we probably need for one of the subproblems to give an acceptable solution and then we would already be done.

3.4 Reducing the search space

When two intervals next to each other is of type 2, the function will be increasing in the variable corresponding to the derivative between them. This means that this variable will be 0 at the optimum, since it has to be non-negative. This is also true when we have a type 2 interval is the first or last interval. This observation can be used to reduce the number of variables needed to be optimized over for certain branches.

Theorem 11. *The optimal solution has at most two intervals of type two next to each other.*

Proof. If a type 2 interval has type 2 intervals on both sides then both v_l and v_r, i.e. the variables corresponding to derivatives, will be 0. In order for such an interval to be acceptable we need $\Delta x < \frac{\Delta t}{3}(v_l + vr - \sqrt{v_l v_r}) = 0$. But since we also have $\Delta x \geq 0$, we cannot get an acceptable solution. This shows that the optimal solution cannot have three intervals of type 2 next to each other. □

Using the same argument one can also say that in the optimal solution the first two intervals cannot be of type 2. The same goes for the last two intervals.

This restriction can be used to decrease the number of different possibilities that needs to be searched in order to find the optimum. The following theorem states approximately how many possibilities are left.

Theorem 12. *The number of subproblems needed to be solved is bounded from above by approximately* 1.84^n *for large* n*, where* n *is the number of intervals.*

Proof. We want to find the number of sequences of length n that can be constructed using the numbers 1 and 2, such that there never are more than two number 2 next to each other. There also cannot be more than one 2 next to the beginning or the end of the sequence.

Let $a_{n,0}, a_{n,1}, a_{n,2}$ be the number of sequences of length n, that can have 0, 1 or 2 twos respectively immediately at the beginning of the sequence. For $a_{n,2}$, if the first number is a 1 we get $a_{n-1,2}$ sequences. If the first number is 2, we get $a_{n-1,1}$ sequences. This gives us $a_{n,2} = a_{n-1,2} + a_{n-1,1}$. With the same reasoning for $a_{n,1}$ and $a_{n,0}$ we get the recursive system of equations

$$\begin{bmatrix} a_{n,2} \\ a_{n,1} \\ a_{n,0} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_{n-1,2} \\ a_{n-1,1} \\ a_{n-1,0} \end{bmatrix}.$$

We also have the initial conditions

$$\begin{bmatrix} a_{1,2} \\ a_{1,1} \\ a_{1,0} \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}.$$

The reason for $a_{1,1} = 1$ is that the "interval" after the last interval also can be seen as type 2, and therefore the last number can only be a 1.

This gives us

$$\begin{bmatrix} a_{n,2} \\ a_{n,1} \\ a_{n,0} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}^{n-1} \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}.$$

We are interested in $a_{n,1}$ since we cannot have two intervals of type two in the beginning.

This equation could be rewritten in a closed form expression, by diagonalizing the matrix. Computing the characteristic equation gives us

$$\lambda^3 - \lambda^2 - \lambda - 1 = 0.$$

This can be solved numerically or symbolically with the help of a computer. There are two complex roots of magnitude approximately 0.73 and a real root at approximatelly 1.84. As n tends to infinity, the complex numbers will go to zero, and the solution will scale with 1.84^n. □

It is not possible in general to say that there cannot be 2 segments of type 2 next to each other. This can be shown by a counterexample. With the dataset $t = [0, 1, 2, 3, 4], \alpha = [0, 1, 1.1, 1.2, 2]$, and $\lambda = 100000$, one gets that the optimum has segments of type [1 2 2 1]. The found function can be seen in figure 1.

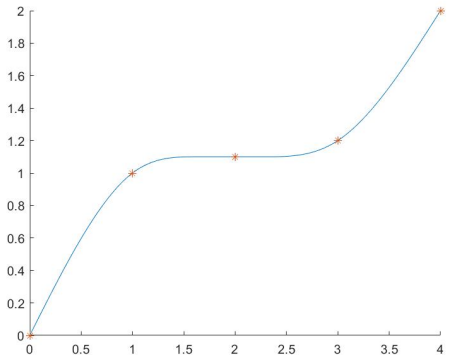


Figure 1: Spline with two type two intervals next to each other.

3.5 Additional implementation details

In the optimization step of the algorithm we use FMINCON in MATLAB, with the interior point method. This is a barrier method and it primarily uses Newtons method, together with a conjugate gradient method if the Newton step does not work [5]. By including the gradient and the Hessian of our objective function the speed was increased significantly.

Since each of the problems are solved numerically we cannot expect to find exactly the minimum in each optimization step. Then if the true minimum is close to the edge of the correct region, then the found minimum may be slightly outside this instead. Therefore it is necessary to have some allowed tolerance when checking if the found solution is accepted. If this is not done it is possible to not find any acceptable solution at all. Letting Δx be 10^{-4} or 10^{-3} seems to work well for most part. There have not been any found cases where this produces several different solutions.

3.6 Solving the original problem directly

Including the gradient and the Hessian in the optimization step greatly increased both speed and performance, also when solving the entire problem rather than one of the subproblems. One possible reason for the increased stability when including the Hessian is that the Hessian is not the same for the two types of functions on the interface between their respective regions. Thus a finite difference approximation

of the Hessian close to the interface is not necessarily a good one. This problem is avoided when including the explicit value of the Hessian.

It is also important to force the function value to infinity whenevever $\Delta x < 0$, otherwise it was possible for Δx to come very close to 0^- and get accepted as inside the correct region even though it was negative.

4 Performance of the algorithm

This section will show some results of the algorithm and discuss its performance on problems of different sizes, as well as for different choices of λ.

4.1 Data sets

Some different data sets were created to test the algorithm. The reason for not using real data sets is that this easily allows for changing the number of points in the data sets, as well as being able to test different shapes.

The data sets are created by choosing some increasing function. The function is shifted in y-direction so that the initial point has the value 0. Some points are chosen on the curve and white noise is added to the function values.

Some shapes that have been tried are linear functions, quadratic functions, the arctan function and different cumulative distribution functions. In figure 2 one can see examples of these.

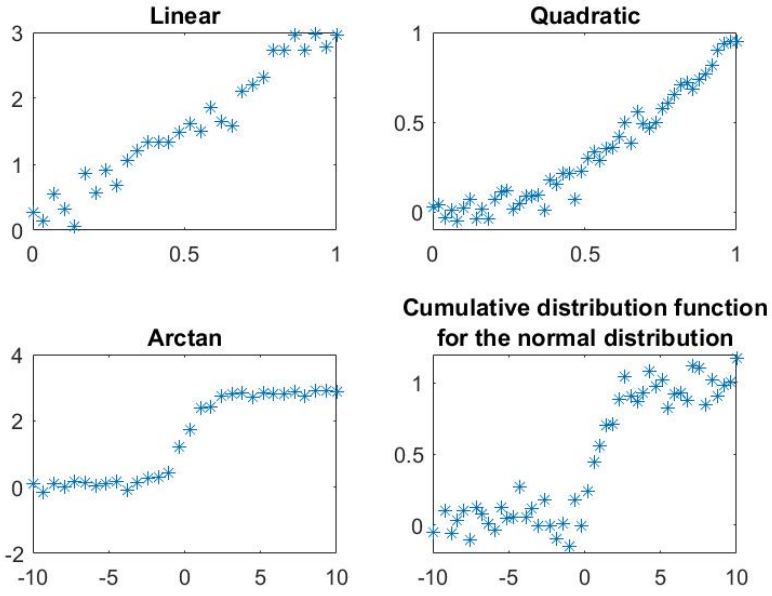


Figure 2: Some data sets that are used.

4.2 Effect of λ

As mentioned before, the size of λ determines how close the curve must fit the data compared to how important it is that it is smoother. When λ goes to zero it becomes more and more important for the curve to have a second derivative of small magnitude. This causes the curve to be closer and closer to a straight line. A straight line will have $v_l = v_r = \frac{\Delta x}{\Delta t}$, and all intervals of type one will be correct. Therefore it is very fast to find the solution when λ is small. In figure 3 one can see the found spline when using a small λ, together with a quadratic data set.

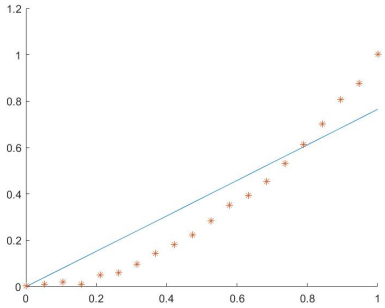


Figure 3: Small λ used to fit a spline to a data set.

On the other hand if λ is large, then the curve needs to be fitted more closely to the points. Without the monotonicity criterion the curve would go almost exactly through the points, with it the curve varies between steep increases and flat sections. An interval will often have $\Delta x < \frac{\Delta t}{3}(v_l + v_r - \sqrt{v_l v_r})$ when the curve is flat but one of the derivatives is large due to the next or previous interval having a large change in x-value. One can see an example of such a curve in figure 4.

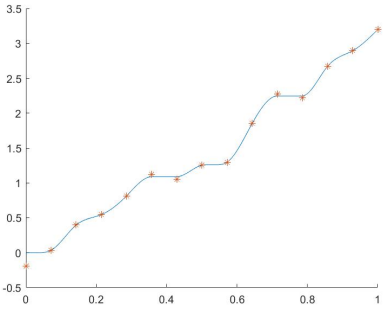


Figure 4: Large λ used to fit a spline to a data set.

As a consequence the solution will have more intervals of type 2 when λ is large. This makes the algorithm significantly slower for these cases.

Usually, neither of these cases are preferable to get a good fit to the points. The optimal is somewhere in between these extreme cases. However, to find this optimal choice of λ one might need to be able to test both of these cases.

4.3 Problem sizes

The time it takes to solve a subproblem varies between approximately 0.02 seconds for a problem with 10 intervals to almost 0.1 seconds when there are 100 intervals. This, of course, depends on the computer being used, but gives a sense of the time needed for the algorithm to run. With 15 intervals or less the algorithm runs fairly fast regardless of shape and choice of λ. On the same computer as above, the program finishes in less than a minute. With 20 intervals the program can run for more than 20 minutes for some data sets with large lambda. When using choices of λ that seem to produce reasonable shapes for the curve, the time it takes depend on the data set. The linear data set generally goes fast and can manage data sets with several hundred points. The same goes for the quadratic data set. An example can be seen in figure 5. Neither of these generally have any intervals of type 2.

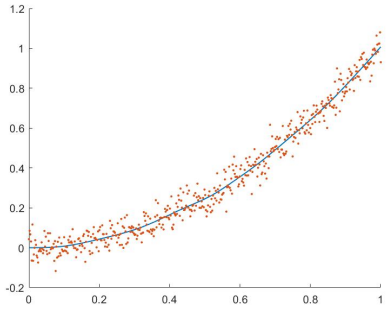


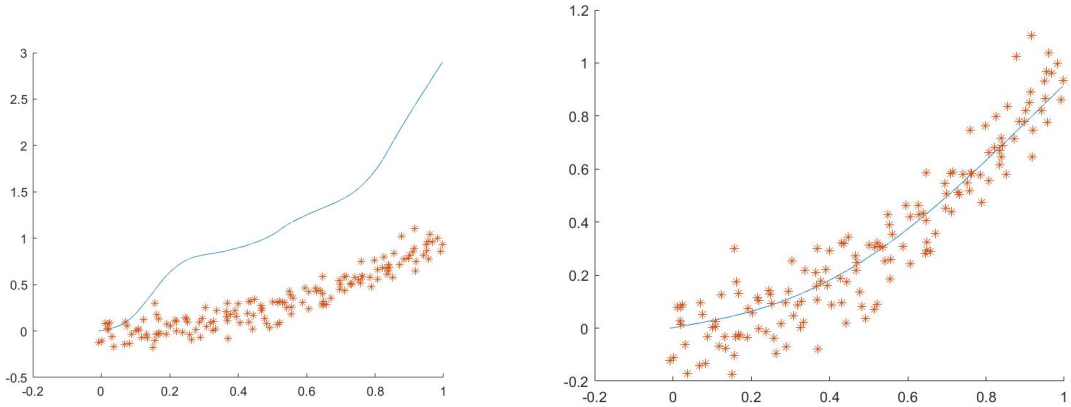
Figure 5: Spline fitted to data set with 500 points.

The arctan data set usually have some intervals that are of type 2 and therefore takes more time, but often it is fast enough for large data sets as well. The tolerance for when a solution is considered accepted also affects. Sometimes it found an acceptable solution when this was set to 10^{-2}, but took too long time when it was set to something smaller such as 10^{-3}. The solution seemed reasonable with the larger tolerance, and it is unclear whether the smaller tolerance would work or not, since it would take a very long time to go through all subproblems.

4.4 Only using the original optimization problem

As mentioned previously, it was possible to improve the performance of solving the original problem rather than the subproblems. When this works it is much faster since the program then only needs to solve one optimization problem. The performance does not seem to depend on λ.

When testing this, most times it works well. Occasionally, the program produce a solution that clearly is not optimal, see figure 6a. However, using a different starting point, in this case the previously bad solution, a much better result was achieved, see figure 6b.



(a) When fitting a spline to the data set a solution that was not optimal was found.

(b) The same data as to the left but with another initial point for the optimization, producing a good result.

Figure 6: Solutions found when solving the original optimization problem

Otherwise this has produced the same solutions as when dividing the problem into subproblems, and can be a good option when it works.

4.5 Some additional observations

It is sometimes possible to guess the correct subproblem by starting with all intervals as type one and solve this problem. By changing the intervals that did not fulfil $\Delta x \geq \frac{\Delta t}{3}(v_l + v_r - \sqrt{v_l v_r})$ to be of the second type instead, quite often one got an acceptable solution.

A possible argument for why this would work for large λ is that then the solution is mostly dependent on the squared distance to the data points. Then it matters less if the interval is of type 1 or 2. Then by changing the intervals that are wrong one

can get a good chance of finding the correct subproblem. On the other hand for small λ it is highly likely that the first subproblem will be acceptable.

Another property that might be possible to make use of is that sometimes changing an interval from type 1 to type 2 the solution to the subproblem becomes much smaller than the optimal value for the correct subproblem. This might indicate that subproblems below this one are unnecessary to test. If this could be proven, it could further reduce the number of problems needed to be solved.

4.6 Discussion and conclusion

We have shown that the problem of finding monotonous smoothing splines can be reduced to a finite dimensional optimization problem and have examined an algorithm for solving this problem.

The number of subproblems needed to be visited in the worst case has been reduced from 2^n to 1.84^n. This is still a large number for relatively small values of data set size n. In practice the speed of the algorithm mainly depends on how many intervals that are of type 2 in the optimal solution. When this number is low, it goes fairly quickly but if not the algorithm takes a lot of time. How many intervals that are of type 2 in the optimal solution depends both on the parameter λ and on the shape of the solution. Generally, choosing a too large value for λ will cause the algorithm to be very slow. This can be a problem when trying different choices of λ to find the optimal choice. Sometimes it is possible to "guess" a subproblem that works, which can be a valid alternative to try. It is also possible that one can find the spline by optimizing the entire problem rather than the subproblems. Still, it seems to often be fast enough close to reasonable values forλ.

It should also be noted that all data sets were created in order to easily vary problem sizes and shapes. Therefore one must keep in mind that the results might be different when using real data sets that might look different from what has been tested.

References

[1] Maad Sasane S., Monotone Smoothing Splines With Bounds, *to appear in Acta Applicandae Mathematicae*, 2020 DOI: 10.1007/s10440-020-00314-0

[2] Wang, Y. *Smoothing Splines, Methods and Applications*, vol. 121 of Monographs on Statistics and Applied Probability. CRC Press, 2011

[3] Egerstedt M., and Martin C., *Control Theoretic Splines*, Princeton University Press, Princeton and Oxford, 2010.

[4] Luenberger, D. G. *Optimization by vector space methods*, John Wiley & Sons, Inc. New York, London, Sydney, Toronto, 1969

[5] Mathworks, https://se.mathworks.com/help/optim/ug/constrained-nonlinear-optimization-algorithms.html#brnpd5f, viewed 27 may 2020

[6] Lars-Christer Böiers, *Mathematical Methods of Optimization*. Studentlitteratur, 2010.

[7] Struwe, M. *Variational Methods*. Springer Verlag, 1996.

[8] Michael Renardy, Robert C. Rogers *An Introduction to Partial Differential Equations* Springer, 2004.

[9] Jürgen Appell, Józef Banas, Nelson José Merentes Díaz, *Bounded variation and around*, De Gruyter, 2014

[10] Evans L. C., *Partial differential equations* American Mathematical Society, cop. 1998

[11] Lars Hörmander, *The analysis of linear partial differential operators. 1, Distribution theory and Fourier analysis*, Springer-Vlg, 1983

Master's Theses in Mathematical Sciences 2020:E38
ISSN 1404-6342

LUTFMA-3418-2020

Mathematics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/