On the complexity of switching linear regression

Fabien Lauer

On the complexity of switching linear regression

Fabien Lauer

Université de Lorraine, CNRS, LORIA, UMR 7503, F-54506 Vandœuvre-lès-Nancy, France

October 23, 2015

Abstract

This technical note extends recent results on the computational complexity of globally minimizing the error of piecewise-affine models to the related problem of minimizing the error of switching linear regression models. In particular, we show that, on the one hand the problem is NP-hard, but on the other hand, it admits a polynomial-time algorithm with respect to the number of data for any fixed data dimension and number of modes.

1 Introduction

Hybrid system identification aims at estimating a model of a system switching between different operating modes from input-output data and is typically setup as a piecewise-affine (PWA) or switching regression problem (see [1, 2] for an overview and a survey of existing approaches). The present paper focuses on the issue of deterministically obtaining a global solution to the switching regression problem. In particular, we are interested in the rather theoretical and yet unanswered question of the existence of an algorithm for this problem with a reasonable time complexity. Therefore, we will concentrate the discussion on computational complexity issues under the classical model of computation known as a Turing machine [3]. In this framework, the time complexity of a problem is the lowest time complexity of an algorithm solving any instance of that problem, where the time complexity of an algorithm is the maximal number of steps occurring in the computation of the corresponding Turing machine program.

Let $\boldsymbol{x}_i \in \mathcal{X} \subseteq \mathbb{Q}^d$ denote the regression vector of index i (possibly built from lagged inputs and outputs of a dynamical system) and $y_i \in \mathbb{Q}$ the corresponding output. Then, we consider the estimation of the parameters $\{\boldsymbol{w}_j\}_{j=1}^n$ of an arbitrarily switching linear model

$$y_i = \boldsymbol{w}_{q_i}^T \boldsymbol{x}_i + v_i,$$

where $q_i \in \mathcal{Q} = \{1, \dots, n\}$ stands for the active mode at index i and $v_i \in \mathbb{Q}$ is a noise term. We assume that the mode q_i is independent of $\boldsymbol{x}_i$, that a data set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N \in (\mathcal{X} \times \mathbb{Q})^N$ of size N significantly larger than the dimension d is available, and that the number of modes n is given. We concentrate on the most common approach minimizing the prediction error over the variables to be estimated, here the classification of the points into modes, i.e., $\boldsymbol{q} \in \mathcal{Q}^N$, and the parameter vectors, $\{\boldsymbol{w}_j\}_{j=1}^n$. Specifically, we formulate the problem in terms of a loss function $\ell : \mathbb{Q} \to \mathbb{Q}^+$, assumed to be computable in polynomial time and to satisfy

$$\begin{cases} \ell(0) = 0, \\ \forall e \in \mathbb{Q},\ \ell(-e) = \ell(e), \\ \forall (e, e') \in \mathbb{Q}^2,\ \ell(e) < \ell(e') \Leftrightarrow |e| < |e'|. \end{cases} \tag{1}$$

Problem 1 (Switching linear regression)**.** *Given a data set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N \in (\mathcal{X} \times \mathbb{Q})^N$ with $\mathcal{X} \subseteq \mathbb{Q}^d$ and an integer $n \in [2, N/d]$, find a global solution to*

$$\min_{\{\boldsymbol{w}_j \in \mathbb{Q}^d\}_{j=1}^n, \boldsymbol{q} \in \mathcal{Q}^N} \ \frac{1}{N} \sum_{i=1}^N \ell(y_i - \boldsymbol{w}_{q_i}^T \boldsymbol{x}_i). \tag{2}$$

Other equivalent formulations based on mixed-integer programming with binary variables encoding $\boldsymbol{q}$ or on continuous optimization can be found [1, 4], together with a number of heuristics subject to local minima [5, 6] or only optimal under specific conditions [7, 8].

Problem 1 can be solved explicitly with respect to (wrt.) $\boldsymbol{q}$ for fixed $\{\boldsymbol{w}_j\}_{j=1}^n$ by assigning each point to the model with minimum error as

$$q_i \in \operatorname*{argmin}_{j\in\{1,\dots,n\}} \ell(y_i - \boldsymbol{w}_j^T \boldsymbol{x}_i), \quad i = 1, \dots, N. \tag{3}$$

Conversely, Problem 1 can be solved wrt. to the $\boldsymbol{w}_j$'s for fixed $\boldsymbol{q}$ as a sequence of linear regression subproblems

$$\min_{\boldsymbol{w}_j\in\mathbb{R}^d} \sum_{\{i:q_i=j\}} \ell(y_i - \boldsymbol{w}_j^T \boldsymbol{x}_i), \quad j = 1, \dots, n. \tag{4}$$

Thus, two global optimization approaches can be readily formulated. The first one tests all possible classifications $\boldsymbol{q}$ and solves the problem wrt. the $\boldsymbol{w}_j$'s for each of them. But, this leads to $n \times n^N$ linear regression subproblems (4) and quickly becomes intractable when N increases. The second approach applies a continuous global optimization strategy to directly estimate $\{\boldsymbol{w}_j\}_{j=1}^n$ under the optimal classification rule (3). However, global optimality cannot be guaranteed without constraints on the $\boldsymbol{w}_j$'s such as box bounds. And even so, the complexity remains exponential in the number of variables nd, for instance for a grid search to obtain a solution with an error that is only guaranteed to be close to the global optimum in finite time.

These straightforward observations illustrate the difficulty of the problem, which is here quantified more formally. In particular, we prove in Sect. 2 that Problem 1 is NP-hard. Nonetheless, we also show in Sect. 3 that the problem can be solved in polynomial time wrt. the number of data for fixed n and d. This result is obtained by generalizing ideas developed in [9] for PWA systems to arbitrarily switched systems, and in particular by deriving for the first time a clear connection between switching regression and linear classification.

2 NP-hardness

In computational complexity, an NP-hard problem is one that is at least as hard as any problem from the class NP of nondeterministic polynomial time decision problems [10]. In particular, NP is the class of all decision problems for which a candidate solution can be certified in polynomial time. Under this definition, we have the following result.

Theorem 1. *With ℓ as in* (1)*, Problem 1 is NP-hard.*

The proof is a direct consequence of the NP-completeness of the following decision form of Problem 1, where an NP-complete problem is one that is both NP-hard and in NP.

Problem 2 (Decision form of switching regression)**.** *Given a data set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N \in (\mathcal{X}\times\mathbb{R})^N$, an integer $n \in [2, N/d]$ and a threshold $\epsilon \geq 0$, decide whether there is a set of vectors $\{\boldsymbol{w}_j \in \mathbb{Q}^d\}_{j=1}^n$ and a labeling $\boldsymbol{q} \in \mathcal{Q}^N$ such that*

$$\frac{1}{N}\sum_{i=1}^{N} \ell(y_i - \boldsymbol{w}_{q_i}^T \boldsymbol{x}_i) \leq \epsilon. \tag{5}$$

We show the completeness of Problem 2 by a reduction from the partition problem, known to be NP-complete [3].

Problem 3 (Partition)**.** *Given a multiset (a set with possibly multiple instances of its elements) of d positive integers, $S = \{s_1, \dots, s_d\}$, decide whether there is a multisubset $S_1 \subset S$ such that*

$$\sum_{s_i\in S_1} s_i = \sum_{s_i\in S\setminus S_1} s_i.$$

Proposition 1. *Problem 2 is NP-complete.*

Proof. Since given a candidate solution $\left(\{\boldsymbol{w}_j\}_{j=1}^n, \boldsymbol{q}\right)$ the condition (5) can be verified in polynomial time, Problem 2 is in NP. Then, the proof of its NP-completeness proceeds by showing that the Partition Problem 3 has an affirmative answer if and only if Problem 2 with $\epsilon = 0$ has an affirmative answer.

Given an instance of Problem 3, set $n = 2$, $N = 2d + 1$ and build a data set such that

$$(\boldsymbol{x}_i, y_i) = \begin{cases} (s_i \boldsymbol{e}_i,\ s_i), & \text{if } 1 \leq i \leq d \\ (s_{i-d} \boldsymbol{e}_{i-d},\ 0), & \text{if } d < i \leq 2d \\ \left(\boldsymbol{s} = \sum_{k=1}^d s_k \boldsymbol{e}_k\ ,\ \frac{1}{2}\sum_{k=1}^d s_k\right), & \text{if } i = 2d+1, \end{cases}$$

where $\boldsymbol{e}_k$ is the kth unit vector of the canonical basis for $\mathbb{Q}^d$. If Problem 3 has an affirmative answer, let I_1 be the set of indexes of the elements of S in S_1 and I_2 the set of indexes of the elements of S not in S_1. Then we can set $\boldsymbol{w}_1 = \sum_{i \in I_1} \boldsymbol{e}_i$ and $\boldsymbol{w}_2 = \sum_{i \in I_2} \boldsymbol{e}_i$, which gives

$$\boldsymbol{w}_1^T \boldsymbol{x}_i = \begin{cases} s_i = y_i, & \text{if } i \leq d \text{ and } i \in I_1 \\ 0, & \text{if } i \leq d \text{ and } i \in I_2 \\ s_{i-d} = y_i, & \text{if } i > d \text{ and } i - d \in I_1 \\ 0, & \text{if } i > d \text{ and } i - d \in I_2 \\ \displaystyle\sum_{k \in I_1} s_k = \frac{1}{2}\sum_{k=1}^d s_k = y_i, & \text{if } i = 2d+1 \end{cases}$$

and

$$\boldsymbol{w}_2^T \boldsymbol{x}_i = \begin{cases} 0, & \text{if } i \leq d \text{ and } i \in I_1 \\ s_i = y_i, & \text{if } i \leq d \text{ and } i \in I_2 \\ 0, & \text{if } i > d \text{ and } i - d \in I_1 \\ s_{i-d} = y_i, & \text{if } i > d \text{ and } i - d \in I_2 \\ \displaystyle\sum_{k \in I_2} s_k = \frac{1}{2}\sum_{k=1}^d s_k = y_i, & \text{if } i = 2d+1. \end{cases}$$

Therefore, for all points, either $\boldsymbol{w}_1^T \boldsymbol{x}_i = y_i$ or $\boldsymbol{w}_2^T \boldsymbol{x}_i = y_i$, and (5) holds with $\epsilon = 0$ and $\boldsymbol{q}$ set as in (3), yielding an affirmative answer for Problem 2.

Assume now that Problem 2 has an affirmative answer with some $\{\boldsymbol{w}_j \in \mathbb{Q}^d\}_{j=1}^n$ and $\epsilon = 0$. Then, the positivity of the loss function implies $\ell(y_i - \boldsymbol{w}_{q_i}^T \boldsymbol{x}_i) = 0$, $i = 1, \ldots, N$, which, by (1) yields

$$\boldsymbol{w}_1^T \boldsymbol{x}_i = y_i \quad \text{or} \quad \boldsymbol{w}_2^T \boldsymbol{x}_i = y_i, \quad i = 1, \ldots, 2d+1. \tag{6}$$

We can always assume that $s_i \neq 0$, since otherwise s_i can be removed from the problem statement. Under this assumption, if $\boldsymbol{w}_1^T \boldsymbol{x}_i = y_i$ for some $i \leq d$, then $w_{1i} = 1$ and $\boldsymbol{w}_1^T \boldsymbol{x}_{d+i} = s_i \neq y_{d+i}$, which further implies $\boldsymbol{w}_2^T \boldsymbol{x}_{d+i} = y_{d+i} = 0$ and $w_{2i} = 0$. Conversely, if $\boldsymbol{w}_2^T \boldsymbol{x}_i = y_i$ for some $i \leq d$, then $w_{2i} = 1$ and $w_{1i} = 0$. Therefore, (6) leads to $w_{1i} \in \{0, 1\}$ and $w_{2i} = 1 - w_{1i}$, $i = 1, \ldots, 2d$. In addition, recall that $y_{2d+1} = \frac{1}{2}\sum_{k=1}^d s_k$, such that, for $i = 2d+1$, (6) yields at least one of the two equalities

$$\boldsymbol{w}_1^T \boldsymbol{x}_{2d+1} = \sum_{k \in \{i \leq d : w_{1i} = 1\}} s_k = \frac{1}{2}\sum_{k=1}^d s_k$$

$$\boldsymbol{w}_2^T \boldsymbol{x}_{2d+1} = \sum_{k \in \{i \leq d : w_{1i} = 0\}} s_k = \frac{1}{2}\sum_{k=1}^d s_k,$$

and a partition corresponding to an affirmative answer for Problem 3 is given by $S_1 = \{s_i : w_{1i} = 1,\ i \leq d\}$. ∎

3 Polynomial time complexity wrt. N

We now turn to the analysis of the computational complexity of Problem 1 wrt. the number of data N, i.e., for fixed n and data dimension d, under the following assumptions.

Assumption 1. *The points $\{\boldsymbol{x}_i\}_{i=1}^N$ are in general position, i.e., no hyperplane of $\mathbb{Q}^d$ contains more than d points. Furthermore, the points $\boldsymbol{z}_i = [\boldsymbol{x}_i^T, y_i]^T$ are also in general position in $\mathbb{Q}^{d+1}$.*

Assumption 2. *Given $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N \in (\mathcal{X} \times \mathbb{Q})^N$, the problem $\min_{\boldsymbol{w}\in\mathbb{Q}^d} \sum_{i=1}^N \ell(y_i - \boldsymbol{w}^T\boldsymbol{x}_i)$ has a polynomial time complexity $T(N)$ for any fixed integer $d \geq 1$.*

Theorem 2. *Under Assumptions 1–2, for given integers d and n, the time complexity of Problem 1 is no more than polynomial in the number of data N and in the order of $T(N)\mathcal{O}(N^{2dn(n-1)})$.*

Theorem 2 is a direct consequence of the existence of an exact algorithm that solves the problem in polynomial time, ensured by Corollary 1 at the end of this section. This algorithm relies on the enumeration of all classifications consistent with (3), which we prove to be in a number polynomial in N below.

We will use recent results on the enumeration of all possible *linear* classifications of a set of N points. In the binary case with two categories, a linear classification is one that can be produced by a separating hyperplane dividing the space in two halfspaces. It is shown in [9] that the number of such hyperplanes producing different classifications of N points is on the order of $\mathcal{O}(N^d)$ in $\mathbb{Q}^d$ and that these hyperplanes can be constructed efficiently. Here, we use an adaptation of these results for linear classifiers, while [9] focused on affine classifiers. This is a minor difference corresponding to the removal of a degree of freedom by forcing the hyperplane to pass through the origin. Since the results of [9] are based on hyperplanes passing through sets of d points, they can be extended in a straightforward manner to deal with linear classifiers by choosing one of these points to be the origin. Thus, we state the following without proof.

Proposition 2 (Adapted from Theorem 3 in [9])**.** *For the class of binary linear classifiers of $\mathcal{X} \subset \mathbb{Q}^d$, $\mathcal{H} = \{h : \mathcal{X} \to \{-1, +1\} : h(\boldsymbol{x}) = sign(\boldsymbol{h}^T\boldsymbol{x}),\ \boldsymbol{h} \in \mathbb{Q}^d\}$, the number of classifications of N points produced by classifiers of $\mathcal{H}$ is bounded for any $N > d$ as*

$$\sup_{S\in\mathcal{X}^N} |\mathcal{H}_S| \leq 2^d \binom{N}{d-1},$$

where $\mathcal{H}_S = \{(h(\boldsymbol{x}_1), \ldots, h(\boldsymbol{x}_N)) \ :\ h \in \mathcal{H}\}$ and $|\cdot|$ denotes the cardinality of a set. In addition, for any set S of N points in general position, there is an algorithm that builds the set $\mathcal{H}_S$ of all linear classifications in $\mathcal{O}\left(2^d \binom{N}{d-1}\right)$ iterations.

In PWA regression, the modes are typically assumed to be linearly separable in the regression space $\mathcal{X}$ and results in the flavor of Proposition 2 can readily be applied to find the optimal classification of the data points [9]. In switching regression, the mode sequence $\{q_i\}$ is arbitrary and we cannot assume the modes to be linearly separable. However, the groups of data pairs $(\boldsymbol{x}_i, y_i)$ associated to different linear models can be "linearly separated" in some sense. More precisely, we will show that the classification rule (3) implicitly entails a combination of two linear classifiers: one applying to the points $\boldsymbol{z}_i = [\boldsymbol{x}_i^T, y_i]^T$ in $\mathbb{Q}^{d+1}$ and another one applying to the regression vectors $\boldsymbol{x}_i$ in $\mathbb{Q}^d$. The equivalence between (3) and these linear classifiers will hold for all points with index not in

$$E = \{i \in \{1, \ldots, N\} :\ \exists (j,k) \in \{1, \ldots, n\}^2,\ j \neq k,\ |y_i - \boldsymbol{w}_j{}^T\boldsymbol{x}_i| = |y_i - \boldsymbol{w}_k{}^T\boldsymbol{x}_i|\}, \tag{7}$$

whose cardinality is bounded by the following lemma.

Lemma 1. *Let E be defined as in (7). Under Assumption 1, $|E| \leq (2d+1)n(n-1)/2$.*

Proof. Let us define, for all (j,k) such that $1 \leq j < k \leq n$, the sets $I_{jk} = \{i : \boldsymbol{w}_j{}^T\boldsymbol{x}_i = \boldsymbol{w}_k{}^T\boldsymbol{x}_i\}$ and $M_{jk} = \{y_i - \boldsymbol{w}_j{}^T\boldsymbol{x}_i = -(y_i - \boldsymbol{w}_k{}^T\boldsymbol{x}_i)\}$. Then, we have $E = \bigcup_{1\leq j<k\leq n} I_{jk} \cup M_{jk}$. Since $\boldsymbol{w}_j{}^T\boldsymbol{x}_i = \boldsymbol{w}_k{}^T\boldsymbol{x}_i \Leftrightarrow (\boldsymbol{w}_j - \boldsymbol{w}_k)^T\boldsymbol{x}_i = 0$, all points $\boldsymbol{x}_i$ with $i \in I_{jk}$ must lie on a hyperplane of

$\mathbb{Q}^d$, and under Assumption 1 we have $|I_{jk}| \leq d$. Similarly, since $y_i - \boldsymbol{w}_j{}^T \boldsymbol{x}_i = -(y_i - \boldsymbol{w}_k{}^T \boldsymbol{x}_i) \Leftrightarrow y_i - (\boldsymbol{w}_j - \boldsymbol{w}_k)^T \boldsymbol{x}_i/2 = 0$, all points $\boldsymbol{z}_i = [\boldsymbol{x}_i^T, y_i]^T$ with $i \in M_{jk}$ must lie on a hyperplane of $\mathbb{Q}^{d+1}$, and Assumption 1 implies that $|M_{jk}| \leq d+1$. Hence, $|E| \leq \sum_{1 \leq j < k \leq n} |I_{jk}| + |M_{jk}| \leq \sum_{1 \leq j < k \leq n} 2d + 1 \leq (2d+1)n(n-1)/2$. □

Proposition 3. *Given a set of parameter vectors $\{\boldsymbol{w}_j\}_{j=1}^n$, let E be defined as in (7). Then, for all $i \notin E$, the classification rule (3) with a loss function satisfying (1) is equivalent to the classification rule*

$$q_i = \operatorname*{argmax}_{j \in \mathcal{Q}} \sum_{k=1}^{j-1} \mathbf{1}_{c_{kj}(\boldsymbol{x}_i, y_i) = -1} + \sum_{k=j+1}^{n} \mathbf{1}_{c_{jk}(\boldsymbol{x}_i, y_i) = +1} \tag{8}$$

implementing a majority vote over a set of $n(n-1)/2$ pairwise classifiers $\{c_{jk}\}_{1 \leq j < k \leq n}$ of $\mathbb{Q}^{d+1}$, where each c_{jk} is a product of binary linear classifiers defined as

$$\forall (\boldsymbol{x}, y) \in \mathbb{Q}^d \times \mathbb{Q},\ c_{jk}(\boldsymbol{x}, y) = g_{jk}(\boldsymbol{z}) h_{jk}(\boldsymbol{x}), \quad 1 \leq j < k \leq n,$$

with linear classifiers respectively operating in $\mathbb{Q}^{d+1}$ and $\mathbb{Q}^d$ as

$$\begin{aligned} \forall \boldsymbol{z} \in \mathbb{Q}^{d+1},\ g_{jk}(\boldsymbol{z}) &= sign([-\overline{\boldsymbol{w}}_{jk}^T,\ 1]^T \boldsymbol{z}), \quad 1 \leq j < k \leq n, \\ \forall \boldsymbol{x} \in \mathbb{Q}^d,\ h_{jk}(\boldsymbol{x}) &= sign(\tilde{\boldsymbol{w}}_{jk}^T \boldsymbol{x}), \quad 1 \leq j < k \leq n, \end{aligned}$$

where $\overline{\boldsymbol{w}}_{jk} = (\boldsymbol{w}_j + \boldsymbol{w}_k)/2$ and $\tilde{\boldsymbol{w}}_{jk} = \boldsymbol{w}_j - \boldsymbol{w}_k$.

Proof. Using the properties of the loss function (1), the classification rule (3) can be rewritten for any $(\boldsymbol{x}_i, y_i)$ with $i \notin E$ as

$$\begin{aligned} & q_i = \operatorname*{argmin}_{k \in \mathcal{Q}} \ell(y_i - \boldsymbol{w}_k^T \boldsymbol{x}_i) \\ \Leftrightarrow \quad & \forall k \in \mathcal{Q} \setminus \{q_i\},\ |y_i - \boldsymbol{w}_{q_i}^T \boldsymbol{x}_i| < |y_i - \boldsymbol{w}_k^T \boldsymbol{x}_i|. \end{aligned} \tag{9}$$

For any triplet $(a, b, y) \in \mathbb{Q}^3$, we have $|y-a| < |y-b|$ if and only if $(a < b \wedge y < (a+b)/2)$ or $(a > b \wedge y > (a+b)/2)$, i.e., $|y-a| < |y-b| \Leftrightarrow (y - (a+b)/2)(a-b) > 0$. Thus, with the notations of Proposition 3, (9) is equivalent to

$$\begin{aligned} & \forall k \in \mathcal{Q} \setminus \{q_i\},\ (y - \overline{\boldsymbol{w}}_{q_i k}^T \boldsymbol{x}_i) \tilde{\boldsymbol{w}}_{q_i k}^T \boldsymbol{x}_i > 0 \\ \Leftrightarrow \quad & \forall k \in \mathcal{Q} \setminus \{q_i\},\ g_{q_i k}(\boldsymbol{z}_i) h_{q_i k}(\boldsymbol{x}_i) = +1 \\ \Leftrightarrow \quad & \sum_{k \in \mathcal{Q} \setminus \{q_i\}} \mathbf{1}_{c_{q_i k}(\boldsymbol{x}_i, y_i) = +1} = n - 1. \end{aligned} \tag{10}$$

In addition, for all $(j,k) \in \{1, \ldots, n\}^2$, $j \neq k$, we have $\overline{\boldsymbol{w}}_{jk} = \overline{\boldsymbol{w}}_{kj}$ and $\tilde{\boldsymbol{w}}_{jk} = -\tilde{\boldsymbol{w}}_{kj}$, so that $c_{jk}(\boldsymbol{x}, y) = -g_{kj}(\boldsymbol{z}) h_{kj}(\boldsymbol{x}) = -c_{kj}(\boldsymbol{x}, y)$. Thus, the classification is entirely determined by a set of $n(n-1)/2$ pairs of linear classifiers g_{jk} and h_{jk}, $1 \leq j < k \leq n$, such that (10) is equivalent to

$$S(q_i) \triangleq \sum_{k=1}^{q_i - 1} \mathbf{1}_{c_{k q_i}(\boldsymbol{x}_i, y_i) = -1} + \sum_{k=q_i+1}^{n} \mathbf{1}_{c_{q_i k}(\boldsymbol{x}_i, y_i) = +1} = n - 1.$$

Given that $\max_{j \in \mathcal{Q}} S(j) \leq n-1$, we obtain that $q_i \in \operatorname{argmax}_{j \in \mathcal{Q}} S(j)$. In addition, for all $q \neq q_i$, $q \in \operatorname{argmax}_{j \in \mathcal{Q}} S(j)$ implies $S(q) = S(q_i) = n-1$, which implies by working the equivalences above backward that $q \in \operatorname{argmin}_{k \in \mathcal{Q}} \ell(y_i - \boldsymbol{w}_k^T \boldsymbol{x}_i)$. However, this is not possible, since for all $i \notin E$, $\operatorname{argmin}_{k \in \mathcal{Q}} \ell(y_i - \boldsymbol{w}_k^T \boldsymbol{x}_i)$ is a singleton. Thus, q_i is the only element in $\operatorname{argmax}_{j \in \mathcal{Q}} S(j)$ as claimed in (8). □

Thanks to Proposition 3, a bound on the number of classifications of N points by (3) can be formed from the product of the number of classifications of the $\boldsymbol{z}_i$'s and of the $\boldsymbol{x}_i$'s.

Theorem 3. *Let us define the set of minimum-error classifications of a data set* $S = \{(\boldsymbol{x}_i, y_i) \in \mathbb{Q}^d \times \mathbb{Q}\}_{i=1}^N$ *as*

$$\mathcal{Q}_S = \left\{ \boldsymbol{q} \in \mathcal{Q}^N : \; q_i = \operatorname*{argmin}_{j \in \mathcal{Q}} \; \ell(y_i - \boldsymbol{w}_j^T \boldsymbol{x}_i), \; i = 1, \ldots, N, \; \boldsymbol{w}_j \in \mathbb{Q}^d, \; j = 1 \ldots, n \right\}.$$

Then, under Assumption 1,

$$\Pi(N) \triangleq \sup_{S \in (\mathbb{Q}^d \times \mathbb{Q})^N} |\mathcal{Q}_S| = \mathcal{O}(N^{2dn(n-1)})$$

and $\mathcal{Q}_S$ *can be computed in* $\mathcal{O}(N^{2dn(n-1)})$ *time.*

Proof. By Proposition 3, any $\boldsymbol{q} \in \mathcal{Q}_S$ must result from a set of $n(n-1)/2$ products of linear classifications, possibly altered for all $i \in E$. Assuming $|E| = l$, n^l such altered classifications can be generated from a "base classification" given by (8). The number of base classifications returned by (8) is bounded by the number of classifications generated by $n(n-1)/2$ pairs of linear classifiers (g_{jk}, h_{jk}), $1 \leq j < k \leq n$, on $N - l$ points. Applying Proposition 2 twice, once in dimension $d+1$ to bound the number of classifications returned by a g_{jk} and once in dimension d to bound the number of classifications returned by a h_{jk}, we can upper bound the number of classifications returned by a product classifier $c_{jk}(\boldsymbol{z}_i) = g_{jk}(\boldsymbol{z}_i) h_{jk}(\boldsymbol{x}_i)$ by $2^{2d+1} \binom{N-l}{d} \binom{N-l}{d-1} = \mathcal{O}(N^{2d-1})$ and obtain an algorithm of a similar time complexity to compute these base classifications. With $p = (2d-1)n(n-1)/2$, this leads to $n^l \mathcal{O}(N^p)$ classifications for each possible set E. Given an upper bound L on $|E|$, summing over all E of cardinality $l \leq L$ yields

$$\begin{aligned} \Pi(N) &= \sum_{l=0}^{L} \binom{N}{l} n^l \mathcal{O}(N^p) = \sum_{l=0}^{L} \mathcal{O}(N^l) n^l \mathcal{O}(N^p) \\ &< L \mathcal{O}(N^L) n^L \mathcal{O}(N^p) = \mathcal{O}(N^{L+p}). \end{aligned}$$

Combining this with the value of the bound L given by Lemma 1 yields the desired result. □

Theorem 3 implies the following for switching regression.

Corollary 1. *Under Assumptions 1–2, there is an algorithm that exactly solves Problem 1 in* $T(N)\mathcal{O}(N^{2dn(n-1)})$ *time.*

Proof. By Theorem 3, an algorithm can generate all classifications that are consistent with (3) and thus visit the optimal classification leading to the minimum of Problem 1 in $\mathcal{O}(N^{2dn(n-1)})$ iterations. Following the discussion of Sect. 1, the algorithm can compute the optimal $\boldsymbol{w}_j$'s for any such classification by solving (4), and thus find the global solution to Problem 1. Under Assumption 2, the cost of solving (4) at each iteration is $nT(N)$, leading to the claimed time complexity. □

Remark 1. *Assumption 1 clearly cannot hold without noise, since, in this case, the points* $\boldsymbol{z}_i$ *precisely belong to the union of n hyperplanes. However, in the noiseless case, an algorithm that runs in time polynomial in N can be devised from the fact that each parameter vector* $\boldsymbol{w}_j$ *can be determined from a subset of d points from the same mode. Assuming that such a subset exists for all modes, it suffices to find these subsets. A straightforward strategy is to consider all subsets of d data points among N for the first mode, d points among* $N - d$ *for the second mode and so on. The number of collections of such disjoint subsets is* $\prod_{k=0}^{n-1} \binom{N-kd}{d} < \binom{N}{d}^n < \left(\frac{N}{d!}\right)^{nd} = \mathcal{O}(N^{nd})$ *and thus polynomial in N. Since testing one of these collections amounts to solving n linear systems in constant time* $\mathcal{O}(nd^3)$*, the resulting algorithm runs in time polynomial in N.*

4 Conclusions

The paper showed that globally minimizing the error of a switching linear regression model is NP-hard, but also that, for fixed data dimension and number of modes, an exact algorithm with polynomial complexity in the number of data exists. This algorithm has an exponential complexity wrt. the data dimension and the number of modes, which strongly limits its practical applicability. Yet, the existence of an algorithm with polynomial complexity in the dimension is unlikely given the NP-hardness of the problem and the fact that it holds also with a fixed number of modes $n = 2$.

References

[1] S. Paoletti, A. L. Juloski, G. Ferrari-Trecate, R. Vidal, Identification of hybrid systems: a tutorial, European Journal of Control 13 (2-3) (2007) 242–262.

[2] A. Garulli, S. Paoletti, A. Vicino, A survey on switched and piecewise affine system identification, in: Proc. of the 16th IFAC Symp. on System Identification (SYSID), 2012, pp. 344–355.

[3] M. Garey, D. Johnson, Computers and Intractability: a Guide to the Theory of NP-Completeness, W.H. Freeman and Company, 1979.

[4] F. Lauer, G. Bloch, R. Vidal, A continuous optimization framework for hybrid system identification, Automatica 47 (3) (2011) 608–613.

[5] F. Lauer, Estimating the probability of success of a simple algorithm for switched linear regression, Nonlinear Analysis: Hybrid Systems 8 (2013) 31–47, supplementary material available at `http://www.loria.fr/~lauer/klinreg/`.

[6] T. Pham Dinh, H. Le Thi, H. Le, F. Lauer, A difference of convex functions algorithm for switched linear regression, IEEE Transactions on Automatic Control 59 (8) (2014) 2277–2282.

[7] R. Vidal, S. Soatto, Y. Ma, S. Sastry, An algebraic geometric approach to the identification of a class of linear hybrid systems, in: Proc. of the 42nd IEEE Conf. on Decision and Control (CDC), Maui, Hawaï, USA, 2003, pp. 167–172.

[8] L. Bako, Identification of switched linear systems via sparse optimization, Automatica 47 (4) (2011) 668–677.

[9] F. Lauer, On the complexity of piecewise affine system identification, Automatica 62 (2015) 148–153, preprint available at `http://hal.archives-ouvertes.fr/hal-01195700/en/`.

[10] V. Blondel, J. Tsitsiklis, A survey of computational complexity results in systems and control, Automatica 36 (9) (2000) 1249–1274.