Global optimization for low-dimensional switching linear regression and bounded-error estimation☆

Fabien Lauer *

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

ABSTRACT

The paper provides global optimization algorithms for two particularly difficult nonconvex problems raised by hybrid system identification: switching linear regression and bounded-error estimation. While most works focus on local optimization heuristics without global optimality guarantees or with guarantees valid only under restrictive conditions, the proposed approach always yields a solution with a certificate of global optimality. This approach relies on a branch-and-bound strategy for which we devise lower bounds that can be efficiently computed. In order to obtain scalable algorithms with respect to the number of data, we directly optimize the model parameters in a continuous optimization setting without involving integer variables. Numerical experiments show that the proposed algorithms offer a higher accuracy than convex relaxations with a reasonable computational burden for hybrid system identification. In addition, we discuss how bounded-error estimation is related to robust estimation in the presence of outliers and exact recovery under sparse noise, for which we also obtain promising numerical results.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

The paper tackles two problems that lie at the core of hybrid dynamical system identification, whose aim is to estimate, from input–output data, a model of a system switching at unknown instants between a number of linear subsystems. More precisely, we consider the minimization of the error of a switching linear model with a fixed number of modes and the iterative maximization of the number of data that can be approximated by a linear model with a bounded error. The latter problem, also known as bounded-error estimation, has an interest outside of hybrid systems as well and in particular for robust estimation in the presence of outliers.

The problems are understood as *global* minimization/maximization problems. However, due to their complexity, most of the literature, as reviewed in Garulli, Paoletti, and Vicino (2012) and Paoletti, Juloski, Ferrari-Trecate, and Vidal (2007), focuses on local optimization or heuristic approaches: for switching regression with a fixed number of modes in Juloski, Weiland, and Heemels (2005), Lauer (2013), Lauer, Bloch, and Vidal (2011), Le,

Lauer, and Bloch (2014), Pham Dinh, Le Thi, Le, and Lauer (2014), Vidal, Soatto, Ma, and Sastry (2003) and for the bounded-error approach to switching regression in Bako (2011), Bemporad, Garulli, Paoletti, and Vicino (2005), Diehm, Maier, Flad, and Hohmann (2013) and Ozay, Sznaier, Lagoa, and Camps (2012). Some of these methods can be proved to yield the global solution but only in specific conditions, such as in the absence of noise for Vidal et al. (2003) and under data-dependent conditions difficult to check in practice for Bako (2011). Recent results showed that, though being NP-hard in general, some hybrid system identification problems, including the minimization of the error of a switching linear model, have a complexity no more than polynomial in the number of data for a fixed data dimension (Lauer, 2015, 2016). However, in practice, the complexity of the corresponding polynomial algorithms remains too high except for small data sets in small dimensions.

Contribution. Global optimization of such difficult problems in general is usually deemed impractical. Hence, we focus on instances where the data can be numerous but should live in a low-dimensional space, as is often the case in a system identification context (most examples in the literature on hybrid system identification have a dimension less than five). In this context, the paper proposes a branch-and-bound approach to the two problems above. Contrary to previous works, such an approach offers unconditional global optimality guarantees, while remaining computationally efficient with large data sets. Branch-and-bound is a standard approach to global optimization, but it was only

considered for hybrid system identification in Roll, Bemporad, and Ljung (2004), where an off-the-shelf solver is applied after a reformulation of the piecewise affine regression problem into a mixed-integer linear or quadratic program, with a number of binary variables proportional to the number of data. At the opposite, the proposed approach can handle larger data sets by developing dedicated optimization algorithms while focusing on the continuous variables of the problems, i.e., the model parameters, rather than the integer variables. Technically, the branch-and-bound approach relies on the derivation of a number of lower bounds on the different cost functions for parameters constrained to lie in a box (a hyperrectangle). In particular, efficiency is obtained thanks to two ingredients: (i) simple lower bounds that can quickly discard boxes with very large costs, and (ii) a constant-classification based criterion that allows us to more tightly lower bound the cost.

Paper organization. Section 2 describes the general branch-and-bound approach adopted to tackle the problems of interest, which are formally described in dedicated sections: Section 3 for switching regression and Section 4 for the bounded-error approach. Then, Section 5 presents numerical results and Section 6 discusses open issues.

Notation. Vectors are written in lowercase bold letters, while matrices are written in uppercase bold letters. For a vector $\boldsymbol{u}$, the kth entry is denoted by u_k, while for a vector $\boldsymbol{u}_j$, its kth entry is $u_{j,k}$. All inequalities between vectors, e.g., $\boldsymbol{u} \leq \boldsymbol{v}$, are meant entrywise. A box $B \subset \mathbb{R}^D$ is a hyperrectangular region of $\mathbb{R}^D$, i.e., $B = [\boldsymbol{u}, \boldsymbol{v}] = \prod_{k=1}^{D}[u_k, v_k]$ with $\boldsymbol{u} \in \mathbb{R}^D$, $\boldsymbol{v} \in \mathbb{R}^D$ such that $\boldsymbol{u} \leq \boldsymbol{v}$. The positive and negative parts of a scalar are denoted by $(\cdot)_+ = \max\{0, \cdot\}$ and $(\cdot)_- = \min\{0, \cdot\}$ and similar notations are used for the corresponding entrywise operations on vectors. Of course, $(\cdot)_+^2$ and $(\cdot)_-^2$ are understood as the squared positive and negative parts of a scalar, i.e., $(\cdot)_+^2 = ((\cdot)_+)^2$ and $(\cdot)_-^2 = ((\cdot)_-)^2$. The notation $|\cdot|$ denotes either the absolute value for real arguments or the cardinality for sets. The indicator function $\mathbf{1}_A$ evaluates to 1 if the Boolean expression A is true and 0 otherwise.

2. General approach

Consider the global minimization of some cost function $J(\boldsymbol{w})$ of a vector of parameters $\boldsymbol{w} \in \mathbb{R}^D$ over a box $B_{\text{init}} = [\boldsymbol{u}_{\text{init}}, \boldsymbol{v}_{\text{init}}] \subset \mathbb{R}^D$, where the different definitions of the cost function J for the problems of interest will be given in dedicated sections below. We attack these problems with a branch-and-bound approach, summarized in Algorithm 1, which takes a data set of regression vectors $\boldsymbol{x}_i \in \mathbb{R}^d$ and target outputs $y_i \in \mathbb{R}$ as inputs. In hybrid system identification, the regression vectors are typically built from lagged inputs and outputs of the system (Paoletti et al., 2007).

The general branch-and-bound scheme relies on computing upper and lower bounds ($\bar{J}$ and $\underline{J}$ in Algorithm 1) on the global optimum $\min_{\boldsymbol{w} \in B_{\text{init}}} J(\boldsymbol{w})$. Then, regions B of the search space in which the local lower bound $\underline{J}(B)$ is larger than the global upper bound $\bar{J}$ can be discarded, reducing the volume left to explore until the relative optimality gap, $(\bar{J} - \underline{J})/\bar{J}$, decreases below a predefined tolerance *TOL*. Here, the considered regions are always boxes, i.e., hyperrectangles. Upper bounds $\bar{J}(B)$ can be easily computed by some local optimization or heuristic method for a problem of interest. Alternatively, $\bar{J}(B)$ can be computed merely as the cost function value at the box base point $\boldsymbol{u}$ or at a random point inside the box, while local optimization is only used periodically. On the other hand, lower bounds $\underline{J}(B)$ require a careful derivation, the efficiency of the approach relying mostly on the tightness of these bounds.

Algorithm 1 retains only the solution yielding the best upper bound $\bar{J} = J(\boldsymbol{w}^*)$. Depending on the value of *TOL*, the algorithm can terminate while there are multiple remaining active boxes possibly

---

Algorithm 1 General branch-and-bound scheme.

Require: A data set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N} \subset \mathbb{R}^d \times \mathbb{R}$, initial box bounds $B_{\text{init}} = [\boldsymbol{u}_{\text{init}}, \boldsymbol{v}_{\text{init}}] \subset \mathbb{R}^D$ and *TOL* > 0. Optionally, an initial guess of $\boldsymbol{w} \in B_{\text{init}}$.
Initialize the global bounds $\underline{J} \leftarrow 0, \bar{J} \leftarrow +\infty$ or $\bar{J} \leftarrow J(\boldsymbol{w})$ if $\boldsymbol{w}$ is provided, and the list of boxes $\mathcal{B} \leftarrow \{B_{\text{init}}\}$.
while $(\bar{J} - \underline{J})/\bar{J} > TOL$ **do**
   Split the current box B into B^1 and B^2 such that $B = B^1 \cup B^2$.
   Compute upper bounds $\bar{J}(B^1)$ and $\bar{J}(B^2)$.
   Update $\bar{J} \leftarrow \min\{\bar{J}, \bar{J}(B^1), \bar{J}(B^2)\}$ and the best solution $\boldsymbol{w}^*$.
   Compute lower bounds $\underline{J}(B^1)$ and $\underline{J}(B^2)$.
   For $k = 1, 2$, append B^k to the list of active boxes $\mathcal{B}$ if $\underline{J}(B^k) \leq \bar{J}$.
   Remove B from the list of active boxes: $\mathcal{B} \leftarrow \mathcal{B} \setminus \{B\}$.
   Select the next box $B \leftarrow \arg\min_{B \in \mathcal{B}} \underline{J}(B)$ and set $\underline{J} \leftarrow \underline{J}(B)$.
end while
return $\boldsymbol{w}^*$ and $\bar{J} = J(\boldsymbol{w}^*) \approx \min_{\boldsymbol{w} \in B_{\text{init}}} J(\boldsymbol{w})$.

---

containing equally good solutions within the tolerance. A possible modification would be to retain a list of solution candidates with cost function values close to the best one rather than a single solution. Since such a modification would be straightforward, in the following, we focus only on the version returning a single solution.

3. Switching linear regression

We consider the identification of a switching system with n modes generating a data set of N points $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \ldots, N$, with

$$y_i = \boldsymbol{w}_{q_i}^T \boldsymbol{x}_i + \xi_i, \tag{1}$$

where $q_i \in \mathcal{Q} = \{1, \ldots, n\}$ is the index of the active mode for the ith point, $\{\boldsymbol{w}_j\}_{j=1}^{n} \subset \mathbb{R}^d$ is a collection of linear model parameter vectors and $\xi_i \in \mathbb{R}$ is a noise term. The aim here is to estimate, from the knowledge of $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ and n only, the concatenated parameter vector $\boldsymbol{w} = [\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_n^T]^T \in \mathbb{R}^{nd}$. Throughout the paper, we assume a similar partitioning of all vectors from $\mathbb{R}^{nd}$, i.e., for $\boldsymbol{u} \in \mathbb{R}^{nd}$, $\boldsymbol{u}_j$ refers to the jth subvector of dimension d in $\boldsymbol{u}$.

Least squares estimates[1] of $\boldsymbol{w}$ and $\boldsymbol{q} = [q_1, \ldots, q_N]^T$ are defined as the global solutions to

$$\min_{\boldsymbol{w} \in \mathbb{R}^{nd}, \boldsymbol{q} \in \mathcal{Q}^N} J_{\text{SWq}}(\boldsymbol{w}, \boldsymbol{q}), \tag{2}$$

with $J_{\text{SWq}}(\boldsymbol{w}, \boldsymbol{q}) = \sum_{i=1}^{N}(y_i - \boldsymbol{w}_{q_i}^T \boldsymbol{x}_i)^2$.

Note that Problem (2) involves N integer variables in $\boldsymbol{q}$, which would imply a worst-case exponential complexity in the number of data for its direct global optimization. Other reformulations based on nN binary variables suffer from a similar limitation, which is why the following considers a continuous optimization point of view.

Using the classification rule[2]

$$q_i(\boldsymbol{w}) = \arg\min_{j \in \mathcal{Q}}(y_i - \boldsymbol{w}_j^T \boldsymbol{x}_i)^2, \quad i = 1, \ldots, N, \tag{3}$$

---

[1] We restrict the presentation to the squared loss function $\ell(e) = e^2$, but similar results could be obtained for instance with the absolute loss $\ell(e) = |e|$.

[2] When the minimum is not unique in (3), ties are arbitrarily broken by returning the minimal index j of the minimum.