Preprints of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

Sparse Identification of Polynomial and Posynomial Models

Giuseppe Carlo Calafiore * **Laurent El Ghaoui** **
Carlo Novara ***

* *Dipartimento di Automatica e Informatica, Politecnico di Torino, Italy. Tel.: +39-011-090.7071; Fax: +39-011-090.7099. E-mail:* giuseppe.calafiore@polito.it
** *EECS and IEOR, UC Berkeley, CA, USA. E-mail:* elghaoui@berkeley.edu
*** *Dipartimento di Automatica e Informatica, Politecnico di Torino, Italy. E-mail:* carlo.novara@polito.it

Abstract: Posynomials are nonnegative combinations of monomials with possibly fractional and both positive and negative exponents. Posynomial models are widely used in various engineering design endeavors, such as circuits, aerospace and structural design, mainly due to the fact that design problems cast in terms of posynomial objectives and constraints can be solved efficiently by means of a convex optimization technique known as geometric programming (GP). However, while quite a vast literature exists on GP-based design, very few contributions can yet be found on the problem of identifying posynomial models from experimental data. Posynomial identification amounts to determining not only the coefficients of the combination, but also the exponents in the monomials, which renders the identification problem numerically hard. In this paper, we outline an approach to the identification of both multivariate polynomial and posynomial models, based on the expansion on a given large-scale basis of monomials. The model is then identified by seeking coefficients of the combination that minimize a mixed objective, composed by a term representing the fitting error and a term inducing sparsity in the representation, which result in a problem formulation of the "square-root LASSO" type, with nonnegativity constraints on the variables. We propose to solve the problem via a sequential coordinate-descent scheme, which is suitable for large-scale implementations.

Keywords: Posynomial models, Identification, Sparse optimization, Square-root LASSO, Coordinate-descent methods.

1. INTRODUCTION

Consider a function $\psi : \mathbb{R}^{n_w} \to \mathbb{R}$ of the form

$$\psi(w) = \sum_{i=1}^{n_c} c_i w^{\alpha_i} \tag{1}$$

where c_i are coefficients, $\alpha_i = [\alpha_{i1} \cdots \alpha_{in_w}] \in \mathbb{R}^{n_w}$ are vectors of exponents, and w^{α_i} is defined as

$$w^{\alpha_i} \doteq \prod_{j=1}^{n_w} w_j^{\alpha_{ij}}.$$

The term $c_i w^{\alpha_i}$ is called a monomial. Two important classes of functions can be cast in the form (1):

(i) polynomials, where $c_i \in \mathbb{R}$, $\alpha_{ij} \in \mathbb{N}_0$, and $w \in \mathbb{R}^{n_w}$;

(ii) posynomials, where $c_i \geq 0$, $\alpha_{ij} \in \mathbb{R}$, and $w \in \mathbb{R}^{n_w}_{++}$ (the positive orthant).

Polynomial models are widely used in many fields of science and technology, and play a fundamental role also in system identification and control, Leontaritis and Billings [1985], Spinelli et al. [2006], Novara [2012]. In system identification, they represent one of the main tools for estimating NARX (non-linear autoregressive with exogenous inputs) systems, see Spinelli et al. [2006], Bonin et al. [2010]; in control, they can be effectively used for the direct design from data of controllers for nonlinear systems, Novara et al. [2013].

Posynomial models are of great importance in many fields as well, ranging from structural design, network flow, optimal control, Beightler and Phillips [1976], Wilde [1978], to aerospace system design, Hoburg and Abbeel [2012], circuit design, Boyd et al. [2005], Daems et al. [2003], antennas, Babakhani et al. [2010] and communication systems, Chiang [2005]. The interest in posynomials is motivated by the fact that they lead to computationally efficient geometric programming models for optimal system design.

Despite a quite consistent number of papers are available in the literature where posynomial models and geometric programming are used for design purposes, very few works can be found that address the key problem of identifying a posynomial model from experimental data; see Daems et al. [2003] for such an exception. The model is in most cases assumed known (i.e., the coefficients c_i and the exponents α_{ij} are assumed known) and then processed by the geometric programming algorithm. On the contrary, in many real-world applications the model is not known a priori and has to be identified from experimental data.

The standard approach to poly/posynomial model identification is to perform an heuristic search finalized at finding a viable model structure, i.e., a suitable set of exponent vectors $\{\alpha_i\}$, see, e.g., Spinelli et al. [2006], Pulecchi and Piroddi [2007], Daems et al. [2003]. Once the exponent vector set has been chosen, the coefficients c_i are estimated by means of convex optimization. A critical issue in this approach is that the model structure search may be extremely time consuming and in most cases leads only to approximate model structures, see Milanese and Novara [2004]. An alternative approach is to assume (or estimate by means of some heuristic) a value $\hat{n}_c$ for the basis cardinality n_c, and then estimate c_i and α_i by means

of nonlinear programming algorithms. However, these kind of algorithms are non-convex and thus do not ensure convergence to the optimal parameter estimate. A third approach, which overcomes the issues of the other two, consists in considering an over-parametrized model and inserting in the optimization problem a sparsity promoting term (or constraint), given by the ℓ_1-norm of the coefficient vector. This term allows one to efficiently select the model structure and, at the same time, to avoid the problem of overfitting. This approach is based on the well-known LASSO (least absolute shrinkage and selection operator) or other similar algorithms, see, e.g., Tibshirani [1996], Kukreja et al. [2006], Bonin et al. [2010], Novara [2012]. The optimization problem is in this case convex but, due to the over-parametrization, it typically involves a very large number of decision variables.

In this paper, we follow this latter approach: we minimize a convex objective, defined as the sum of a regularized accuracy term based on the ℓ_2-norm of the estimation residual, and a sparsity-inducing term given by a weighted ℓ_1-norm of the coefficient vector. We name this approach *regularized square-root LASSO* or rsqrt-LASSO, since it is similar to LASSO but presents three differences which may give advantages in terms of computational efficiency and model regularity. The first one is to use in the objective function an accuracy objective that is the square-root of the one used in LASSO. Thanks to this feature, we obtain an a-priori sufficient condition for a monomial appearing in the over-parameterization to be null. This condition (called *feature elimination* condition, El Ghaoui et al. [2012]) can be verified very efficiently, and can thus be used in a pre-optimization phase to eliminate all the monomials which have very low relevance in explaining the data. The second difference is to include an ℓ_2 regularization in the accuracy term, allowing us to account for uncertainty in the data and to improve the numerical conditioning. The third difference consists in using a weighted ℓ_1-norm of the coefficient vector in place of the standard ℓ_1-norm. This allows for more generality in problems where the entries of c have different scales. Along with the basic rsqrt-LASSO model, we also consider a nonnegative version of the problem (named nonnegative rsqrt-LASSO), where variables are constrained to be nonnegative, as required for the identification of posynomials.

In order to solve the rsqrt-LASSO and nnrsqrt-LASSO problems, we propose a large-scale-capable iterative algorithm based on sequential coordinate descent, which is able to deal with problems involving a large number of decision variables.

2. IDENTIFICATION OF POLY/POSYNOMIALS

Consider a poly/posynomial

$$\psi^o(w) = \sum_{i=1}^{n_c} c_i^o w^{\alpha_i^o} \tag{2}$$

where the coefficients c_i^o, the exponent vectors α_i^o and the expansion cardinality n_c are not known. Suppose that a set of noise-corrupted measurements is available:

$$\mathfrak{D} = \{y(k), w(k)\}_{k=1}^m,$$

where $y(k) = \psi^o(w(k)) + e(k)$, and $e(k) \in \mathbb{R}$ is a noise term. The problem considered in the paper is to estimate from these data the unknown parameters c_i^o, α_i^o, $i = 1, \dots, n_c$, and the cardinality n_c. To solve this problem, we define an over-parametrized poly/posynomial family

$$\psi(w) = \sum_{i=1}^{n} x_i w^{\alpha_i} \tag{3}$$

where $n \gg n_c$. In real-world situations, this over-parametrization can be obtained from the available prior information on the exponents α_{ij}^o. For example, a certain exponent may be unknown but it can be known to be an integer in a given interval; another one may be known to be fractional in another interval; another one can be known to be negative, etc.

More formally, suppose that the following prior information is available on the exponents: $\alpha_{ij} \in Q_j$, where Q_j is a set of exponents which, on the basis of the available prior information, can be considered reasonable for the variable w_j. Then, the set of exponent vectors defining the over-parametrization (3) can be constructed as $S_\alpha \doteq \{\alpha_i\}_{i=1}^n = \prod_{j=1}^{n_w} Q_j$, where $\prod$ denotes the Cartesian product. Note that this approach can be adopted also if an exponent is known to belong to a continuous (finite) interval, in which case the set Q_j can be obtained by properly discretizing the interval. If the information is correct, then S_α is guaranteed to contain the true exponent vectors: $S_\alpha \supset S_{\alpha^o} \doteq \{\alpha_i^o\}_{i=1}^{n_c}$.

2.1 A square-root LASSO formulation

Model identification is here performed by minimizing with respect to the coefficients x_i an objective function defined as the sum of an accuracy objective and a sparsity-promoting term, allowing us to select, in the over-parametrized family, a parsimonious model structure. Define $y = [y(1) \cdots y(m)]^\top$, $x = [x_1 \cdots x_n]^\top$, and

$$\Phi = \begin{bmatrix} w(1)^{\alpha_1} & \cdots & w(1)^{\alpha_{n_w}} \\ \vdots & \ddots & \vdots \\ w(m)^{\alpha_1} & \cdots & w(m)^{\alpha_{n_w}} \end{bmatrix}.$$

The objective we consider is of the form

$$f(x) \doteq \left\| \begin{bmatrix} \Phi x - y \\ \sigma x \end{bmatrix} \right\|_2 + \lambda^\top |x|, \tag{4}$$

where $\sigma \geq 0$, $\lambda \in \mathbb{R}^n$ with $\lambda \geq 0$, and $|x|$ denotes a vector whose entries are the absolute values of the entries in x. We define, for notational compactness,

$$\tilde{\Phi} \doteq \begin{bmatrix} \Phi \\ \sigma I \end{bmatrix}, \quad \tilde{y} \doteq \begin{bmatrix} y \\ 0 \end{bmatrix}, \quad \tilde{\phi}_i \doteq \begin{bmatrix} \phi_i \\ \sigma e_i \end{bmatrix},$$

where $\tilde{\phi}_i$, $i = 1, \dots, n$, denotes the i-th column of $\tilde{\Phi}$, and e_i is the i-th vector of the standard basis of $\mathbb{R}^n$.

Note that $\lambda^\top |x|$ is a weighted ℓ_1-norm. Vector λ is thus a penalty factor which quantifies the tradeoff between the accuracy objective $\|\tilde{\Phi}x - \tilde{y}\|_2$ and the term $\lambda^\top |x|$, which is a proxy for sparsity in the solution, see, e.g., Fuchs [2005], Tropp [2006], Donoho et al. [2006], Candes and Tao [2006]. Clearly, for $\lambda = \gamma \mathbf{1}$ (where $\mathbf{1}$ is a vector with all entries equal to one), and $\sigma = 0$, the rsqrt-LASSO problem coincides with the standard sqrt-LASSO. The use of the sparsity promoting term $\lambda^\top |x|$ instead of the standard term $\gamma \|x\|_1$ allows for more generality, in problems where the entries of x have different scales. The regularization parameter $\sigma \geq 0$ is introduced to improve the numerical conditioning of the problem, guaranteeing (if $\sigma > 0$) that $\tilde{\Phi}$ has full rank, and that the ℓ_2 term remains differentiable for all x.

We hence consider the following two optimization problems, which we name regularized square-root LASSO (rsqrt-LASSO)

$$p^* \doteq \min_{x \in \mathbb{R}^n} f(x), \tag{5}$$

and nonnegative regularized square-root LASSO (nnrsqrt-LASSO)

$$p_+^* \doteq \min_{x \in \mathbb{R}_+^n} f(x), \tag{6}$$

where $\mathbb{R}^n_+ \doteq \{x \in \mathbb{R}^n : x \geq 0\}$ (the inequality is element-wise). The first is to be used for polynomial model identification, and the second for posynomial model identification.

Remark 1. Notice that the cardinality n of the set S_α, and hence the dimension of the decision vector x, may be very large, since it is given by the product of the cardinalities of Q_j, for $j = 1, \ldots, n_w$. For this reason, although the two previous problems are standard convex optimization problems, they may not be practically solved using standard interior-point methods for convex optimization. Actually, in some cases, even just storing in memory the data matrix Φ may be unfeasible due to dimensionality issues.

In the following sections, we describe a simple scheme for solving both the unconstrained and the constrained versions of the regularized sqrt-LASSO problem, based on a two-phase procedure. In the first phase, we apply a feature elimination step to eliminate a-priori all variables that are guaranteed to be zero at optimum, thus possibly reducing the dimensionality of the problem. In the second phase, we apply a coordinate-descent scheme to the reduced problem, in order to find the optimal solution. This latter phase is based on the fact that we can find in "closed form" an optimal solution to the univariate restriction of the above problems.

We shall assume throughout that $y \neq 0$, since for $y = 0$ the optimal solution of both problems (5), (6) is trivially $x^* = 0$.

3. DUAL FORMULATIONS AND FEATURE ELIMINATION

We next state dual formulations of the rsqrt-LASSO and nnrsqrt-LASSO problems, and then show how a feature elimination condition is obtained from these dual formulations.

3.1 Dual of the rsqrt-LASSO problem

The dual of problem (5) can be expressed in the following form (derivations are omitted here for space reasons; see Calafiore et al. [2014] for full details)

$$\begin{aligned} p^* = \max_u \quad & -u^\top \tilde{y} && (7)\\ \text{s.t.:} \quad & \|u\|_2 \leq 1 \\ & |\tilde{\phi}_i^\top u| \leq \lambda_i, i = 1, \ldots, n. && (8) \end{aligned}$$

3.2 Dual of the nnrsqrt-LASSO problem

The dual of the nnrsqrt-LASSO problem (6) can be expressed in the following form (see again Calafiore et al. [2014] for a derivation)

$$\begin{aligned} p^*_+ = \max_u \quad & -u^\top \tilde{y} && (9)\\ \text{s.t.:} \quad & \|u\|_2 \leq 1 \\ & \tilde{\phi}_i^\top u + \lambda_i \geq 0, i = 1, \ldots, n. && (10) \end{aligned}$$

3.3 Safe feature elimination

We next analyze the dual formulations of problems (5), (6) in order to derive a simple sufficient condition that permits to predict when an entry x_i is zero at optimum, and hence to eliminate a priori some features (i.e., columns of $\tilde{\Phi}$) from the problem. This type of condition, first introduced by El Ghaoui et al. [2012] in the context of the standard LASSO problem, is named *safe feature elimination*. Observe that

$$\max_{\|u\|_2 \leq 1} |\tilde{\phi}_i^\top u| = \|\tilde{\phi}_i\|_2 = \left\| \begin{bmatrix} \phi_i \\ \sigma e_i \end{bmatrix} \right\|_2 .$$

Therefore, if for some $i \in \{1, \ldots, n\}$ it holds that

$$\left\| \begin{bmatrix} \phi_i \\ \sigma e_i \end{bmatrix} \right\|_2^2 = \|\phi_i\|_2^2 + \sigma^2 < \lambda_i^2$$

then the corresponding constraint in (8), as well as in (10), will certainly be satisfied with strict inequality, that is, it will be *inactive* at the optimum. This means that it can be safely eliminated from the dual optimization problem, without changing the optimal objective value. Defining

$$\mathcal{F}(\lambda) \doteq \{i : \|\phi_i\|_2^2 + \sigma^2 \geq \lambda_i^2, \ i = 1, \ldots, n\},$$

we thus have that

$$\begin{aligned} p^* = \max_u \quad & -u^\top \tilde{y} && (11)\\ \text{s.t.:} \quad & \|u\|_2 \leq 1 \\ & |\tilde{\phi}_i^\top u| \leq \lambda_i, i \in \mathcal{F}(\lambda), \end{aligned}$$

which is the dual of the "reduced" primal problem

$$p^* = \min_\xi \|\tilde{\Phi}_{\mathcal{F}(\lambda)}\xi - \tilde{y}\|_2 + \lambda^\top |\xi|, \qquad (12)$$

where $\tilde{\Phi}_{\mathcal{F}(\lambda)}$ is a matrix containing by columns vectors $\tilde{\phi}_i$, $i \in \mathcal{F}(\lambda)$, and ξ is a decision variable vector, having dimension equal to the cardinality of $\mathcal{F}(\lambda)$. In other words, the features x_i in the primal problem (5) corresponding to indexes i in the set $\mathcal{E}(\lambda)$ complementary to $\mathcal{F}(\lambda)$

$$\mathcal{E}(\lambda) \doteq \{i : \|\phi_i\|_2^2 + \sigma^2 < \lambda_i^2, \ i = 1, \ldots, n\},$$

are certainly zero at the optimum, that is

$$\|\phi_i\|_2^2 + \sigma^2 < \lambda_i^2 \quad \Rightarrow \quad x_i^* = 0. \qquad (13)$$

Similarly, we have that

$$\begin{aligned} p^*_+ = \max_u \quad & -u^\top \tilde{y} && (14)\\ \text{s.t.:} \quad & \|u\|_2 \leq 1 \\ & \tilde{\phi}_i^\top u + \lambda_i \geq 0, i \in \mathcal{F}(\lambda), \end{aligned}$$

is the dual of the "reduced" primal problem

$$p^*_+ = \min_{\xi \geq 0} \|\tilde{\Phi}_{\mathcal{F}(\lambda)}\xi - \tilde{y}\|_2 + \lambda^\top |\xi|. \qquad (15)$$

3.4 When is $x = 0$ optimal?

Point $x = 0$ is optimal for problem (5) if and only if $p^* = \|\tilde{y}\|_2$, which is equivalent to $u = -\tilde{y}/\|\tilde{y}\|_2$ being optimal (hence feasible) for the dual problem. This happens if and only if

$$|\tilde{\phi}_i^\top \tilde{y}| \leq \lambda_i \|\tilde{y}\|_2, \quad i = 1, \ldots, n,$$

that is, since $\tilde{\phi}_i^\top \tilde{y} = \phi_i^\top y$, $\|\tilde{y}\|_2 = \|y\|_2$, if an only if

$$|\phi_i^\top y| \leq \lambda_i \|y\|_2, \quad i = 1, \ldots, n.$$

Similarly, point $x = 0$ is optimal for problem (6) if and only if $p^*_+ = \|\tilde{y}\|_2$, which is equivalent to $u = -\tilde{y}/\|\tilde{y}\|_2$ being optimal (hence feasible) for the dual problem, which happens if and only if

$$\tilde{\phi}_i^\top \tilde{y} \leq \lambda_i \|\tilde{y}\|_2, \quad i = 1, \ldots, n,$$

or, equivalently,

$$\phi_i^\top y \leq \lambda_i \|y\|_2, \quad i = 1, \ldots, n.$$

4. UNIVARIATE SOLUTION OF RSQRT-LASSO

Consider the following rsqrt-LASSO problem with a single scalar variable x

$$\min_{x\in\mathbb{R}} f(x) \doteq \left\| \begin{bmatrix} \phi x - y \\ \sigma e x - \xi \end{bmatrix} \right\|_2 + \lambda|x|,$$

where $\lambda, \sigma \geq 0$, $\phi \in \mathbb{R}^m, y \in \mathbb{R}^m$, $\xi \in \mathbb{R}^n$ are given, and e is a vector of all zeros, except for an entry in generic position i, which is equal to one, and correspondingly we postulate that $\xi_i = 0$, thus it holds that $e^\top \xi = 0$. We set for convenience

$$\tilde{\phi} \doteq \begin{bmatrix} \phi \\ \sigma e \end{bmatrix}, \quad \tilde{y} \doteq \begin{bmatrix} y \\ \xi \end{bmatrix}, \tag{16}$$

thus the problem rewrites to

$$\min_{x\in\mathbb{R}} f(x) \doteq \|\tilde{\phi}x - \tilde{y}\|_2 + \lambda|x|, \tag{17}$$

We assume that $\tilde{y} \neq 0$ and $\tilde{\phi} \neq 0$, otherwise the optimal solution is simply $x = 0$. Let us define

$$x_{\text{ls}} \doteq \frac{\tilde{\phi}^\top \tilde{y}}{\|\tilde{\phi}\|_2^2} = \frac{\phi^\top y}{\|\phi\|_2^2 + \sigma^2},$$

which corresponds to the solution of the problem for $\lambda = 0$. The following proposition holds, see Calafiore et al. [2014] for a proof.

Proposition 1. Consider problem (17), with $\tilde{y} \neq 0$, $\tilde{\phi} \neq 0$, $\lambda \geq 0$.

(1) $x^* = 0$ is an optimal solution for (17) if and only if
$$|\tilde{\phi}^\top \tilde{y}| \leq \lambda \|\tilde{y}\|_2$$
(notice, in particular, that if $\|\tilde{\phi}\|_2 \leq \lambda$, then the above condition is certainly satisfied, hence $x^* = 0$).

(2) If $|\tilde{\phi}^\top \tilde{y}| > \lambda \|\tilde{y}\|_2$ (hence $\|\tilde{\phi}\|_2 > \lambda$), then the optimal solution of (17) is given by
$$x^* = x_{\text{ls}} - \text{sgn}\,(x_{\text{ls}}) \frac{\lambda}{\|\tilde{\phi}\|_2^2} \sqrt{\frac{\|\tilde{\phi}\|_2^2 \|\tilde{y}\|_2^2 - (\tilde{\phi}^\top \tilde{y})^2}{\|\tilde{\phi}\|_2^2 - \lambda^2}}. \tag{18}$$

4.1 Univariate solution of nnrsqrt-LASSO

The solution of the univariate nnrsqrt-LASSO problem in scalar variable x

$$\min_{x\geq 0} f(x) \doteq \|\tilde{\phi}x - \tilde{y}\|_2 + \lambda|x|, \tag{19}$$

can be readily obtained from the solution of the corresponding unconstrained problem (17), by the following reasoning. Since (19) is a convex optimization problem in one variable and one linear inequality constraint, its optimal solution is either on the boundary of the feasible set (in this case, at $x = 0$), or it coincides with the solution of the unconstrained version of the problem. Thus, we solve the unconstrained problem (17): if this solution is nonnegative, then it is also the optimal solution to (19); if it is negative, then the optimal solution to (19) is $x = 0$. Since the sign of the solution of (17) is simply the sign of $\tilde{\phi}^\top \tilde{y}$, we can state the following proposition.

Proposition 2. Consider problem (19), with $\tilde{y} \neq 0$, $\tilde{\phi} \neq 0$, $\lambda \geq 0$.

(1) $x^* = 0$ is an optimal solution for (19) if and only if
$$\tilde{\phi}^\top \tilde{y} \leq \lambda \|\tilde{y}\|_2.$$

(2) Otherwise, the optimal solution of (19) is given by
$$x^* = x_{\text{ls}} - \frac{\lambda}{\|\tilde{\phi}\|_2^2} \sqrt{\frac{\|\tilde{\phi}\|_2^2 \|\tilde{y}\|_2^2 - (\tilde{\phi}^\top \tilde{y})^2}{\|\tilde{\phi}\|_2^2 - \lambda^2}}. \tag{20}$$

Remark 2. For the specific structure of $\tilde{\phi}$, $\tilde{y}$ in (16), we have that

$$\|\tilde{\phi}\|_2^2 = \|\phi\|_2^2 + \sigma^2, \quad \tilde{\phi}^\top \tilde{y} = \phi^\top y, \quad \|\tilde{y}\|_2^2 = \|y\|_2^2 + \|\xi\|_2^2,$$

and the solutions in Proposition 1 and Proposition 2 can be expressed accordingly in terms of $\phi^\top y$, $\|\phi\|_2$, $\|y\|_2$, $\|\xi\|_2$, and σ, λ. In particular, the condition for $x = 0$ being optimal becomes

$$|\phi^\top y| \leq \lambda \sqrt{\|y\|_2^2 + \|\xi\|_2^2},$$

which, in particular, is satisfied if $\|\phi\|_2^2 + \sigma^2 \leq \lambda^2$.

Notice further that $\tilde{\phi}x - \tilde{y} \neq 0$ for $x = 0$, since we assumed $\tilde{y} \neq 0$, and that, for $\sigma > 0$, $\tilde{\phi}x - \tilde{y} \neq 0$ also for $x \neq 0$, since the i-th entry of ξ is zero by definition. Therefore, for $\sigma > 0$, the ℓ_2-norm part of the objective is always nonzero, and hence differentiable.

5. SEQUENTIAL COORDINATE DESCENT SCHEME

We next outline a sequential coordinate-descent scheme for the rsqrt-LASSO problem (5). Suppose all variables x_j, $j \in \{1, \ldots, n\} \setminus i$, are fixed to some numerical values, and we wish to minimize the objective in (5) with respect to the scalar variable x_i. We have that

$$\begin{aligned} f_i(x_i) &\doteq \|\sum_{j=1}^n \tilde{\phi}_j x_j - \tilde{y}\|_2 + \sum_{j=1}^n \lambda_j |x_j| \\ &= \|\tilde{\phi}_i x_i - \tilde{y}(i)\|_2 + \lambda_i |x_i| + \sum_{j\neq i} \lambda_j |x_j|, \end{aligned}$$

where we defined $\tilde{y}(i) \doteq \tilde{y} - \sum_{j\neq i} \tilde{\phi}_j x_j$. We thus have that

$$x_i^* \doteq \arg\min_{x_i} f_i(x_i) = \arg\min_{x_i} \|\tilde{\phi}_i x_i - \tilde{y}(i)\|_2 + \lambda_i |x_i|,$$

where the minimizer x_i^* is readily computed by applying Proposition 1.

A sequential coordinate-descent scheme works by updating the variables x_i sequentially, according to the above univariate minimization criterion. The scheme of the algorithm is as follows.

(1) Initialize $x^{(0)} = 0$ (n-vector of zeros), $k = 1$;
(2) For $i = 1, \ldots, n$, let
$$x_i^{(k)} = \arg\min_{x_i} f(x_1^{(k)}, \ldots, x_{i-1}^{(k)}, x_i, x_{i+1}^{(k-1)}, \ldots, x_n^{(k-1)});$$
(3) If stopping criterion is met, finish and return $x^{(k)}$, else set $k \leftarrow k + 1$, and goto 2.

Remark 3. As a stopping criterion, one may use a standard check on sufficient progress in objective reduction, or the approach described in Section 5.1, based on the evaluation of a lower bound on the duality gap.

Remark 4. Observe that, due to Proposition 1, all variables x_i for which $\|\tilde{\phi}_i\|_2 \leq \lambda_i$ are *never* updated by the algorithm, i.e., they remain fixed at their initial zero value. The inner loop on i can thus be sped up by considering only the indices i such that $\|\tilde{\phi}_i\|_2 > \lambda_i$, which can be determined a priori (feature elimination).

Remark 5. The same coordinate-descent scheme can be used also for solving the nnrsqrt-LASSO problem (6), by using the result in Proposition 2 for updating the i-th coordinate.

Remark 6. The function $f(x)$ in (4) that we minimize using coordinate descent is convex and composite:

$$f(x) = f_0(x) + \sum_{i=1}^n \psi_i(x_i),$$

where ψ_i are convex and nonsmooth. In the unconstrained case, we have $\psi_i(x_i) = \lambda_i|x_i|$. The constrained case, where $x_i \geq 0$, can also be tackled as an unconstrained one, by considering $\psi_i(x_i) = \lambda_i|x_i| + I_+(x_i)$, where $I_+(x_i)$ is equal to zero if $x_i \geq 0$ and it is $+\infty$ otherwise.

Further, function $f_0(x) = \|\tilde{\Phi}x - \tilde{y}\|_2$ is convex and, for $\sigma > 0$ and $y \neq 0$, it is differentiable over all $x \in \mathbb{R}^n$. In this situation, the sequential coordinate descent algorithm is guaranteed to converge to an optimal point on both the rsqrt-LASSO and the nnrsqrt-LASSO problems; see, e.g., Theorem 5.1 in Tseng [2001].

5.1 Dual-bound based stopping criterion

Inspecting the primal and dual problems (5), (7), we see that if x^* is primal optimal, then the dual-optimal variable u must be

$$u^* = \frac{\tilde{\Phi}x^* - \tilde{y}}{\|\tilde{\Phi}x^* - \tilde{y}\|_2}.$$

This suggests considering, for the candidate solution $x^{(k)}$ at iteration k of the algorithm, an associated vector

$$u^{(k)} \doteq \alpha^{(k)}\tilde{u}^{(k)}, \quad \tilde{u}^{(k)} \doteq \frac{\tilde{\Phi}x^{(k)} - \tilde{y}}{\|\tilde{\Phi}x^{(k)} - \tilde{y}\|_2},$$

where

$$\alpha^{(k)} = \begin{cases} 1 & \text{if } |\tilde{\Phi}^\top \tilde{u}^{(k)}| \leq \lambda \\ \min\limits_i \dfrac{\lambda_i}{|\tilde{\phi}_i^\top \tilde{u}^{(k)}|} & \text{otherwise.} \end{cases}$$

Such $u^{(k)}$ is, by construction, feasible for the dual problem (7), hence

$$d^{(k)} \doteq -\tilde{y}^\top u^{(k)} = \alpha^{(k)} \frac{\|\tilde{y}\|_2^2 - \tilde{y}^\top \tilde{\Phi}x^{(k)}}{\|\tilde{\Phi}x^{(k)} - \tilde{y}\|_2}$$

is a lower bound on the primal optimal value p^*, that is $d^{(k)} \leq p^* \leq p^{(k)}$, where $p^{(k)} \doteq f(x^{(k)})$. Hence, if at iteration k it holds that $p^{(k)} - d^{(k)} \leq \epsilon$, we can terminate the algorithm with a solution $x^{(k)}$ that guarantees ϵ-suboptimality.

An analogous approach can be followed for determining a dual lower bound for the nnrsqrt-LASSO problem (6). The only difference is in the choice of $\alpha^{(k)}$, which is now given by

$$\alpha^{(k)} = \begin{cases} 1 & \text{if } \tilde{\Phi}^\top \tilde{u}^{(k)} \geq -\lambda \\ \min\limits_{\{i:\, \tilde{\phi}_i^\top \tilde{u}^{(k)} < -\lambda_i\}} \dfrac{\lambda_i}{|\tilde{\phi}_i^\top \tilde{u}^{(k)}|} & \text{otherwise.} \end{cases}$$

6. A NUMERICAL TEST

As a numerical experiment, we considered the problem of identifying a posynomial model for the drag force (per unit length) of a NACA 4412 airfoil. This force is evaluated as a function of the air flow density ρ, the wing chord η, the incidence angle θ and the flow velocity v, that is

$$F_D = \psi^o(w)$$

where $w = [\rho\ \eta\ \theta\ v]^\top$. The values $\psi^o(w)$ are obtained via simulations based on CFD (computational fluid dynamics), by integration of the Navier-Stokes equations. Each evaluation is numerically very costly, thus it is of interest to obtain a simple model for F_D, to be used, for instance, in a later stage of system evaluation or design. In this example, we identified a posynomial model for the drag force of the airfoil, from data obtained from the CFD simulations. The posynomial form is of interest since it allows the application of geometric programming algorithms, which in turn allow for efficient optimization of the airfoil characteristics, see, e.g., Hoburg and Abbeel [2012].

PARAM.	Minimum	Maximum	Dimension
ρ	0.039	1.2250	[kg/m^3]
η	0.1	1	[m]
θ	-5	10	[deg]
v	0	40	[m/s]

Table 1. Parameter intervals considered in the CFD simulations.

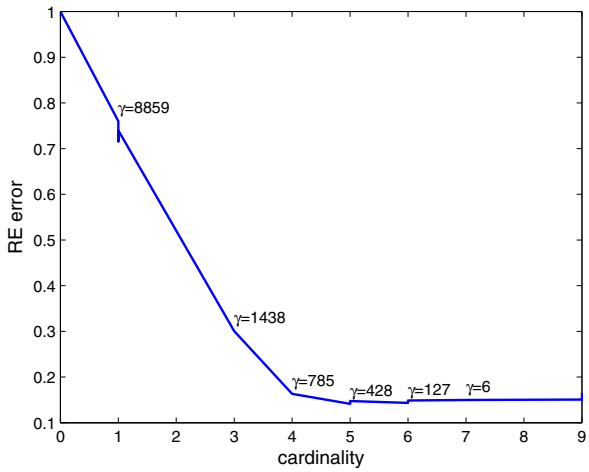

Figure 1. Pareto trade-off curve.

A set $\mathfrak{D} = \{y(k) = \psi^o(w(k)), w(k)\}_{k=1}^m$ of $m = 50$ input-output data points has been obtained, for randomly chosen values of ρ, η, θ and v in the intervals shown in Table 1. The exponent sets

$$Q_j = \{-2, -1, 0, 1, 2\},\ j = 1, \ldots, 4. \tag{21}$$

have been assumed, following the approach described in Section 2. This choice has been made after a preliminary trial and error process. Sets Q_j with exponents ranging from -3 to 3 taking non integer values have been also considered in this process but no significant improvements in terms of model accuracy have been observed. For $m = 50$ and for the exponent sets (21), Φ results to be a 50×625 matrix.

We set for simplicity $\lambda = \gamma\mathbf{1}$, $\sigma = \gamma/10$, and we considered several values of γ, logarithmically spaced in the interval $[1, 10^5]$. For each value of γ, the optimization problem (6) has been solved using the approach described in Sections 3-5. For each value of γ, the following quantities have been recorded:

- the cardinality (i.e., the number of nonzero entries) of the solution x of the optimization problem (6);
- the relative error $RE = \|\Phi x - y\|_2 / \|y\|_2$.

Figure 1 shows the Pareto trade-off curve, reporting the RE values versus the solution cardinality. Based on this curve, the parameter value $\gamma = 785$ has been chosen, since providing the best trade-off between the model complexity (measured by the cardinality of x) and its accuracy (measured by the relative error RE).

In order to verify the reliability of an identified model, we carried out a leave-one-out (LOO) cross validation, on a subset of the available data. In particular, we used for cross validation data points $w(j)$ that lie within 0.75% from the boundary of the the hyperrectangle defining the minimum and maximum deviation for each parameter (as defined in Table 1). This was done to avoid points near the boundary

of the w domain, which are too close to the non-explored region.

For each pair $(y(j), w(j))$ in the LOO validation set, a posynomial model has been identified from the data set $\mathfrak{D} \setminus (y(j), w(j))$. This model has then been tested on the single datum $(y(j), w(j))$, and the relative error $\nu_j = |y(j) - \hat{y}(j)|/\|y_{LOO}\|_2$ has been evaluated, where $\hat{y}(j)$ is the output provided by the model, and $\|y_{LOO}\|_2$ is the Euclidean norm of the vector with entries $y(j)$, for j in the validation set. The accumulated relative error is given by $AE = \sqrt{\sum_j \nu_j^2}$. In our experiment, with $\gamma = 785$, we obtained $AE = 0.25$. This value appears to be quite low: a model identified using the proposed approach is able to approximate the unknown function quite accurately, even if only 50 points are used to explore its 4-dimensional domain.

The same LOO validation has been performed considering $\gamma = 1438$ and $\gamma = 127$, obtaining $AE = 0.38$ and $AE = 0.25$, respectively. The model identified using $\gamma = 785$ has thus the most advantageous trade-off between complexity and accuracy. This model is given by

$$\psi(w) = x_{340}\eta v^2 + x_{440}\rho v^2 + x_{465}\rho\eta v^2 + x_{565}\rho^2 v^2$$

where $x_{340} = 1.2746 \times 10^{-4}$, $x_{440} = 3.5469 \times 10^{-3}$, $x_{465} = 2.8703 \times 10^{-4}$, and $x_{565} = 5.0722 \times 10^{-4}$ (the units of these coefficient can be inferred from Table 1). It is interesting to note that a dependence of the drag force on the square velocity has been found by the algorithm and this result is consistent with the well-known drag equation. No significant dependence on the incidence angle θ has been observed. A possible interpretation for this latter result is that the range considered for θ is not sufficiently large compared to the ranges considered for ρ, η and v (see Table 1) and, consequently, the force variations due to θ are negligible with respect to those produced by the other three parameters.

We next discuss a few relevant aspects related to the identification process. The safe feature elimination discussed in Section 3.3, reduced the number of columns of Φ from 625 to 222 (this latter is the average value obtained in the LOO validation), suggesting that this elimination phase can be quite useful in practical large-scale problems. The time taken for applying the safe elimination and solving the optimization problem (6) with the approach described in Sections 3-5 is about 0.35 seconds on a PC with a Core i7 processor and a RAM memory of 8GB (average time obtained in the LOO validation).

Acknowledgments: We thank Valentina Dolci (Politecnico di Torino) for providing us with the fluid dynamic simulation data used in the example.

REFERENCES

A. Babakhani, J. Lavaei, J. Doyle, and A. Hajimiri. Finding globally optimum solutions in antenna optimization problems. In *IEEE International Symposium on Antennas and Propagation*, 2010.

C.S. Beightler and D.T. Phillips. *Applied geometric programming*. Wiley, New York, 1976.

M. Bonin, V. Seghezza, and L. Piroddi. NARX model selection based on simulation error minimisation and LASSO. *IET Control Theory and Applications*, 4(7): 1157–1168, 2010. doi: 10.1049/iet-cta.2009.0217.

S.P. Boyd, S.J. Kim, D.D. Patil, and M.A. Horowitz. Digital circuit optimization via geometric programming. *Operation Research*, 53(6):899–932, 2005.

G.C. Calafiore, L. El Ghaoui, and C. Novara. Sparse identification of posynomial models. *arXiv:1311.4362*, 2014.

E.J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52 (12):5406 –5425, dec. 2006. ISSN 0018-9448. doi: 10.1109/TIT.2006.885507.

M. Chiang. Geometric programming for communication systems. *Commun. Inf. Theory*, 2:1–154, 2005.

W. Daems, G. Gielen, and W. Sansen. Simulation-based generation of posynomial performance models for the sizing of analog integrated circuits. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 22(5):517–534, 2003.

D.L. Donoho, M. Elad, and V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6 – 18, jan. 2006. ISSN 0018-9448. doi: 10.1109/TIT.2005.860430.

L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination for the LASSO and sparse supervised learning problems. *Pacific Journal of Optimization*, 8(4):667–698, 2012.

J.J. Fuchs. Recovery of exact sparse representations in the presence of bounded noise. *IEEE Transactions on Information Theory*, 51(10):3601 –3608, oct. 2005. ISSN 0018-9448. doi: 10.1109/TIT.2005.855614.

W. Hoburg and P. Abbeel. Geometric programming for aircraft design optimization. In *8th AIAA MDO Specialist Conference*, Honolulu, HI, USA, 2012.

S.L. Kukreja, J. Lofberg, and M.J. Brenner. A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification. In *14th IFAC Symp. on System Identification*, pages 814–819, Newcastle, Australia, 2006.

I.J. Leontaritis and S.A. Billings. Input-output parametric models for non-linear systems - part I: deterministic non-linear systems. *Int. J. Control*, 41:303–328, 1985.

M. Milanese and C. Novara. Set membership identification of nonlinear systems. *Automatica*, 40/6:957–975, 2004.

C. Novara. Sparse identification of nonlinear functions and parametric set membership optimality analysis. *IEEE Transactions on Automatic Control*, 57(12):3236–3241, 2012. doi: 10.1109/TAC.2012.2202051.

C. Novara, L. Fagiano, and M. Milanese. Direct feedback control design for nonlinear systems. *Automatica*, 49(4): 849–860, 2013.

T. Pulecchi and L. Piroddi. A cluster selection approach to polynomial NARX identification. In *American Control Conference*, pages 852–857, New York City, USA, 2007.

W. Spinelli, L. Piroddi, and M. Lovera. A two-stage algorithm for structure identfication of polynomial NARX models. In *American Control Conference*, pages 2387–2392, 2006.

R Tibshirani. Regression shrinkage and selection via the Lasso. *Royal. Statist. Soc B.*, 58(1):267–288, 1996.

J.A. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030 –1051, mar. 2006. ISSN 0018-9448. doi: 10.1109/TIT.2005.864420.

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. of Optimization Theory and Applications*, 109(3):475–494, 2001.

D. Wilde. *Globally optimal design*. Wiley interscience publication, 1978.