HAL
open science

Sparse signal reconstruction for nonlinear models via piecewise rational optimization

Arthur Marmin, Marc Castella, Jean-Christophe Pesquet, Laurent Duval

To cite this version:

Arthur Marmin, Marc Castella, Jean-Christophe Pesquet, Laurent Duval. Sparse signal reconstruction for nonlinear models via piecewise rational optimization. Signal Processing, 2021, 179, pp.107835:1-107835:13. 10.1016/j.sigpro.2020.107835 . hal-02972442v1

HAL Id: hal-02972442

https://hal.science/hal-02972442v1

Submitted on 20 Oct 2020 (v1), last revised 4 Dec 2020 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Sparse Signal Reconstruction for Nonlinear Models via Piecewise Rational Optimization

Arthur Marmin[a,*], Marc Castella[b], Jean-Christophe Pesquet[a], Laurent Duval[c]

[a] *University Paris-Saclay, CentraleSupélec Center for Visual Computing, Inria, 9 Rue Joliot Curie, 91190 Gif-sur-Yvette, France*
[b] *SAMOVAR, CNRS, Télécom SudParis, Institut Polytechnique de Paris, 91011 Evry Cedex, France*
[c] *ESIEE Paris, University Paris-Est, LIGM, Noisy-le-Grand, and IFP Energies nouvelles, Rueil-Malmaison, France.*

Abstract

We propose a method to reconstruct sparse signals degraded by a nonlinear distortion and acquired at a limited sampling rate. Our method formulates the reconstruction problem as a nonconvex minimization of the sum of a data fitting term and a penalization term. In contrast with most previous works which settle for approximated local solutions, we seek for a global solution to the obtained challenging nonconvex problem. Our global approach relies on the so-called Lasserre relaxation of polynomial optimization.

We here specifically include in our approach the case of piecewise rational functions, which makes it possible to address a wide class of nonconvex exact and continuous relaxations of the ℓ_0 penalization function. Additionally, we study the complexity of the optimization problem. It is shown how to use the structure of the problem to lighten the computational burden efficiently. Finally, numerical simulations illustrate the benefits of our method in terms of both global optimality and signal reconstruction.

Keywords: polynomial and rational optimization, global optimization, ℓ_0 penalization, sparse modelling

1. Introduction

Sparse signals, i.e. signals composed of a few spikes, are of particular interest. They either occur naturally in many areas or emerge after sparsifying

Preliminary versions of this work were presented in [1] and in [2]

*Corresponding author

Email addresses: `arthur.marmin@centralesupelec.fr` (Arthur Marmin), `marc.castella@telecom-sudparis.eu` (Marc Castella), `jean-christophe.pesquet@centralesupelec.fr` (Jean-Christophe Pesquet), `laurent.duval@ifpen.fr` (Laurent Duval)

transformations such as time-frequency or wavelet decompositions [3, 4]. However, accurate data acquisition of sparse signals from real-world measurements remains an open challenge. The difficulty of the problem is further increased when acquiring data at a reduced rate. This is however an important practical situation, since it permits faster acquisitions for high-throughput experiments and analysis.

A common approach to recover the original signal from the observations is first to define a well-chosen criterion and then to minimize it. The criterion is often composed of two terms: a fit function depending on the investigated model as well as the observations, and a (possibly composite) regularization term that allows good estimates to be selected among those consistent with the data [5]. However, few methods today are able to deal with nonlinear models and to globally optimize sparsity promoting criteria. Indeed, integrating any of these two properties in the criterion often yields an intricate optimization problem that is difficult to solve.

Thence, to deal with nonlinear effects, linearization techniques are often used since the vast majority of available methods only apply to linear models [6, 7, 8] or to models with weaker linearity assumptions [9, 10, 11]. On the other hand, the standard approach to promote sparse solutions consists in adding an ℓ_0 penalization to a data-fit cost function which leads to NP hard optimization problems [12, 13]. Consequently, several surrogates to the ℓ_0 penalization have been suggested, the simplest one being the ℓ_1 norm. The latter has the enjoyable property of being convex, which simplifies the optimization task [14, 15], but it also strongly penalizes high values of the variables and thus introduces a bias in the solutions. Albeit providing good results, the nonconvex Geman-McClure function [16] also tends to introduce bias. Therefore further relaxations of ℓ_0 function have been investigated [17]. A major drawback is that those relaxations are nonconvex and result in optimization problems which are difficult to solve globally in the sense that currently available algorithms only converge to local solutions and therefore may be highly dependent on their initialization [17, 18, 19, 20, 21, 7, 22].

In the case of a linear model, a first approach for ensuring global convergence of an exact relaxation of the ℓ_0 function has been proposed in [13] and is based on mixed-integer programming. This work proposes a different approach grounded on the global minimization of the broad class of piecewise rational functions under polynomial constraints. Based on it, we propose a novel recovery method for sparse signals from subsampled observations obtained through a noisy model involving nonlinear functions. More precisely, we show that the fit function and the regularization term can be modeled as piecewise rational functions. Fortunately, many well-known good approximations to the ℓ_0 penalization satisfy the latter property [23, 17, 24, 25, 26, 8, 2]. Moreover, various nonlinear degradations, such as saturation, can be modeled with rational functions. Hence, several criteria of interest for reconstructing sparse signals which have been nonlinearly degraded can be modeled, or faithfully approximated, as piecewise rational. We then reformulate the corresponding piecewise rational optimization problem as the minimization of a sum of rational functions, for

which the recent framework of Lasserre's hierarchy [27] can be applied. This framework relaxes a polynomial optimization problem into a hierarchy of convex semi-definite programming (SDP) problems whose solutions converge to a global solution to the initial polynomial problem. SDP problems are playing an important role in our methodology, however, solving large dimensional SDP problems remains nowadays an open challenge. Therefore, we study the overall complexity of the SDP relaxations and show how to reduce it efficiently in several ways. We especially emphasize the benefit of subsampling. Our contribution is twofold:

- First, we investigate a wide range of continuous approximations to the ℓ_0 penalty and we extend the framework of Lasserre's hierarchy to piecewise rational functions in order to minimize the resulting nonconvex criterion. Unlike standard approaches, we are able to establish theoretical guarantees on the global optimum of the original optimization problem. In particular, we provide a unified view of our previous works [1, 28, 2]. The framework and the nonlinear observation model have first been proposed in [28] while the subsampling has been introduced in [1]. However, in both [28, 1], the regularizer was restricted to a Geman-Mcclure potential. We propose here to use a much richer class of regularizers which was introduced in [2] but only for a simple linear model.

- Second, through a complexity analysis and extensive simulations, we show how the structure of the problem and subsampling allow us to alleviate the computational burden of the original Lasserre's framework. Our approach can be successfully applied to signal processing and compressed sensing problems as illustrated by the provided example inspired by the acquisition of signals in gas chromatography.

Our article is organized as follows: Section 2 introduces our model and criterion while Section 3 presents the class of approximations to the ℓ_0 penalty we consider before reformulating the minimization of our criterion as a rational optimization problem. Section 4 details how to solve such optimization problem by leveraging its inherent structure. Section 5 first studies the complexity of the obtained SDP problems before explaining how to decrease it efficiently. Section 6 presents numerical simulations in order to validate our method. Section 7 concludes our work.

We introduce the following notation: $*$ is the convolution operator, for any nonnegative integers n and k, $\mathbb{S}^n$ (resp. $\mathbb{S}^n_+$) is the set of $n \times n$ real symmetric (resp. symmetric positive semi-definite) matrices, $\binom{n}{k}$ is the binomial coefficient "among n choose k", $\lfloor \cdot \rfloor$ (resp. $\lceil \cdot \rceil$) is the greatest (resp. smallest) integer lower (resp. greater) than its argument, $|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_n$ denotes the absolute value of a multi-index $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ of size n, and $\mathbb{N}^n_t$ is the subset of multi-indices whose absolute value is less than or equal to t. The superscript $^\top$ indicates the transpose of a matrix. For a given set $\mathcal{X}$, $\mathbb{1}_{\{\cdot \in \mathcal{X}\}}$ is the characteristic function of $\mathcal{X}$ with $\mathbb{1}_{\{x \in \mathcal{X}\}} = 1$ if x is in $\mathcal{X}$ and 0 otherwise. For a given polynomial p,

we define the following operator

$$\mathrm{d}_p = \left\lceil \frac{\operatorname{degree} p}{2} \right\rceil , \tag{1}$$

and we denote by $\mathbf{p}$ a vector composed of the coefficients corresponding to monomials in p up to the total degree of p.

2. Observation and signal model

2.1. Our observation model

We consider the reconstruction of an unknown discrete-time sparse signal $\overline{\mathbf{x}}$ of length T. The measurement process deteriorates $\overline{\mathbf{x}}$ in the following way: the peaks it contains are enlarged and the sensors introduce a saturation effect. As common in the literature, these degradations are modeled respectively by a convolution with a finite impulse response filter and by a memoryless nonlinear function Φ. The filter coefficients are given by a vector $\mathbf{h}$ of length L. Finally, a noise is superimposed, which is modeled by an additive vector term $\mathbf{w}$ with samples drawn from an i.i.d. zero-mean Gaussian distribution.

An important feature of our model is its ability to deal with subsampling of the measured signal during the acquisition. As in many applications such as chromatography and spectroscopy, the physical limitations may allow only subsampled data acquisition, we introduce a decimation operator D. Interestingly, we will see that our approach is applicable in this context and allows one to use well-suited penalization terms to promote sparsity. Defining the observation vector $\mathbf{y}$ of size U after subsampling, the corresponding modeling equation finally reads

$$\mathbf{y} = D\big(\Phi(\mathbf{h} * \overline{\mathbf{x}}) + \mathbf{w}\big) . \tag{2}$$

Model (2) can emulate narrow-peak signals from gas chromatography experiments [29, 30]. In this case, the filter $\mathbf{h}$ has a discretized Gaussian shape. This choice arises from traditional stochastic or plate modeling, representing a Galton-Hennequin bell distribution [31, Chapter 3]. Peak saturation is also modeled, which cannot be done in standard practice in analytical chemistry. According to [32], filter lengths L from 3 to 9 samples may suffice for a relatively accurate estimation of the peak area, a quantity related to the concentration of a particular molecule.

We will be interested in regular decimation patterns D_δ where all the elements indexed with a multiple of an integer δ are deleted, namely

$$D_\delta\big((s_t)_{t\in[\![1,T]\!]}\big) = \big(s_{\Delta(u,\delta)}\big)_{u\in[\![1,U]\!]} , \tag{3}$$

where

$$U = T - \lfloor T/\delta \rfloor \tag{4}$$

and Δ is defined as

$$(\forall u \in [\![1,U]\!]) \quad \Delta(u,\delta) = u + \left\lfloor \frac{u-1}{\delta-1} \right\rfloor .$$

We denote by D_∞ the identity operator that preserves the entire signal. Let us illustrate the two decimation patterns D_2 and D_4 on the example vector $\mathbf{s} = [s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8]^\top$

$$\mathbf{s} \xmapsto{D_2} [s_1, s_3, s_5, s_7]^\top = \left(s_{\Delta(u,2)}\right)_{u\in[\![1,4]\!]}$$
$$\mathbf{s} \xmapsto{D_4} [s_1, s_2, s_3, s_5, s_6, s_7]^\top = \left(s_{\Delta(u,4)}\right)_{u\in[\![1,6]\!]} .$$

The smaller parameter δ, the higher the decimation and harder the reconstruction of the signal $\overline{\mathbf{x}}$.

To estimate the original signal $\overline{\mathbf{x}}$, we minimize a penalized criterion $\mathcal{J}$ composed of two terms:

$$(\forall \mathbf{x} \in \mathbb{R}^T) \quad \mathcal{J}(\mathbf{x}) = f_{\mathbf{y}}(\mathbf{x}) + \mathcal{R}_\lambda(\mathbf{x}) . \tag{5}$$

The first one $f_{\mathbf{y}}$ is a fit measure with respect to the acquired measurements $\mathbf{y}$ while the second one $\mathcal{R}_\lambda$ is a regularization term which will be discussed next in more detail in Section 2.2.

As a fit function, we choose the standard least-squares error between $\mathbf{y}$ and the output of the noiseless model for a given estimate $\mathbf{x}$ of the original signal $\overline{\mathbf{x}}$

$$(\forall \mathbf{x} \in \mathbb{R}^T) \quad f_{\mathbf{y}}(\mathbf{x}) = \|\mathbf{y} - D_\delta(\Phi(\mathbf{h} * \mathbf{x}))\|_2^2 = \|\mathbf{y} - D_\delta(\Phi(\mathbf{H}\mathbf{x}))\|_2^2 ,$$

where $\mathbf{H}$ is a Toeplitz band matrix corresponding to the convolution with $\mathbf{h}$. Because of the transformation Φ, the fit function $f_{\mathbf{y}}$ is possibly nonconvex. This is in contrast with more classical linear models in which the fit function reduces to the quadratic function $\mathbf{x} \longmapsto \|\mathbf{y} - D_\delta(\mathbf{H}\mathbf{x})\|_2^2$. In our approach, other fit functions $f_{\mathbf{y}}$ can be chosen to model different problems as long as they are rational. In the following, the nonlinear function Φ is assumed to be rational and to act component-wise. Setting the components of $\mathbf{x}$ with nonpositive index to be identically zero in order to unclutter notation, $f_{\mathbf{y}}$ hence reads as a sum of rational functions

$$f_{\mathbf{y}}(\mathbf{x}) = \sum_{u=1}^{U} \underbrace{\left(y_u - \Phi\Big(\sum_{l=1}^{L} h_l x_{\Delta(u,\delta)-l+1} \Big) \right)^2}_{g_u\left(x_{\Delta(u,\delta)-L+1}, \ldots, x_{\Delta(u,\delta)}\right)} ,$$

where $(g_u)_{u\in[\![1,U]\!]}$ are rational functions in L variables.

2.2. Properties of the original signal and examples of ℓ_0 approximations

The unknown original signal $\overline{\mathbf{x}}$ sought by the reconstruction method is assumed to be sparse. In other words, it comprises only few peaks and many of its components are zero. Following this assumption, the second term $\mathcal{R}_\lambda$ in (5) is a sparsity-promoting penalization weighted by a positive parameter λ. Ideally, we would like $\mathcal{R}_\lambda$ to be the sparsity measure $\lambda\ell_0$ (where ℓ_0 counts the number of

nonzero elements) but, in order to derive computationally efficient optimization techniques, a suitable separable approximation is substituted for it, which reads

$$(\forall \mathbf{x} = (x_t)_{t\in[\![1,T]\!]} \in \mathbb{R}^T) \quad \mathcal{R}_\lambda(\mathbf{x}) = \sum_{t=1}^{T} \Psi_\lambda(x_t)\,. \tag{6}$$

Common approaches consist in using either convex functions Ψ_λ such as the ℓ_1 norm, or nonconvex ones that still maintain the convexity of the overall criterion [22]. However, a good approximation $\Psi_\lambda : \mathbb{R} \to \mathbb{R}$ to the ℓ_0 function requires the following three properties [17] leading to nonconvex criteria: unbiasedness for large values, sparsity to reduce the complexity of the model by setting small values to zero, and continuity to ensure the stability of the model. In contrast with [28, 1] where the Geman-McClure nonconvex ℓ_0 approximation was used and introduced bias in the estimate, we propose here a much wider class of piecewise rational function approximations that satisfy the three mentioned properties. Those approximations extend significantly our previous work to settings of more practical interest.

Several examples of functions Ψ_λ shown in the literature to yield good approximations to the ℓ_0 function are actually piecewise rational functions, for which we will show in this article that exact minimization is achievable. We list below examples of the most commonly used piecewise rational approximations to the ℓ_0 penalization that appear in several areas such as imaging or statistics. Figure 1 displays the graph of those functions on $[-3, 3]$.

- Capped ℓ_p [23, 25, 26]:

$$\Psi_\lambda(x) = |x|^p \, \mathbb{1}_{\{|x|\leq\lambda\}} + \lambda^p \mathbb{1}_{\{|x|>\lambda\}}\,.$$

- Smoothly clipped absolute deviation (SCAD) [17]: ($\gamma \in]2, +\infty[$)

$$\begin{aligned}\Psi_\lambda(x) = \lambda\,|x|\,\mathbb{1}_{\{|x|\leq\lambda\}} + \frac{(\gamma+1)\lambda^2}{2}\mathbb{1}_{\{|x|>\gamma\lambda\}} \\ - \frac{\lambda^2 - 2\gamma\lambda\,|x| + x^2}{2(\gamma-1)}\mathbb{1}_{\{\lambda<|x|\leq\gamma\lambda\}}\,,\end{aligned}$$

- Minimax concave penalty (MCP) [24]: ($\gamma \in \mathbb{R}_+^*$)

$$\Psi_\lambda(x) = \left(\lambda\,|x| - \frac{x^2}{2\gamma}\right)\mathbb{1}_{\{|x|\leq\gamma\lambda\}} + \frac{\gamma\lambda^2}{2}\mathbb{1}_{\{|x|>\gamma\lambda\}}\,,$$

- Continuous exact ℓ_0 (CEL0) [8]: ($\gamma \in \mathbb{R}_+^*$)

$$\Psi_\lambda(x) = \lambda - \frac{\gamma^2}{2}\left(|x| - \frac{\sqrt{2\lambda}}{\gamma}\right)^2 \mathbb{1}_{\left\{|x|\leq\frac{\sqrt{2\lambda}}{\gamma}\right\}}\,.$$

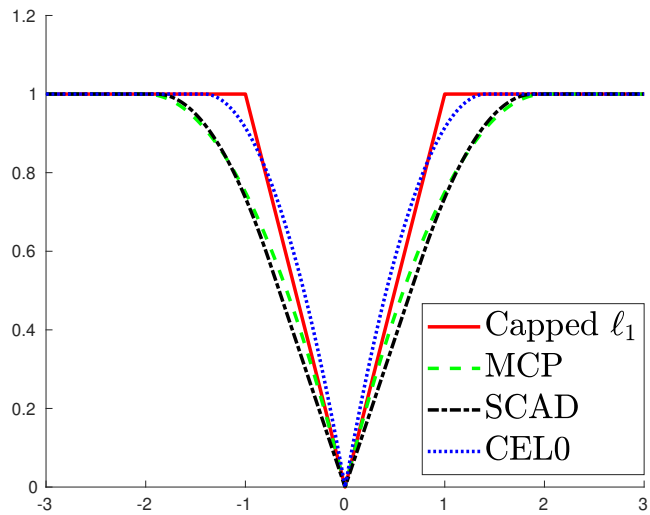


Figure 1: Examples of continuous relaxation of ℓ_0 penalization ($\lambda = 1$, $\gamma_{\text{SCAD}} = 2.5$, $\gamma_{\text{MCP}} = 2$, $\gamma_{\text{CEL0}} = 1$).

Although CEL0 and MCP share a similar expression for the function Ψ_λ, they are quite different in their overall form $\mathcal{R}_\lambda$ due to the choice of the parameter γ. In (6), this parameter for MCP is fixed for all the samples x_t, while for CEL0, its value is adapted to each sample. In the above penalization, the lower the parameter γ, the tighter the approximation to the ℓ_0 penalization but the stronger also the nonconvexity. An important remark concerning the above examples is that, when Φ is set to the identity, a suitable choice of the parameter γ guarantees that the global minimizers of the criterion $f_{\mathbf{y}} + \mathcal{R}_\lambda$ are exactly the global minimizers of the criterion $f_{\mathbf{y}} + \lambda\ell_0$ [33]. The choice of γ depends on the parameter λ and the norm of the columns of $D_\delta\mathbf{H}$. This behavior provides important insights and guarantees on the quality of the above functions as penalization terms to enforce sparsity of the solutions.

3. Rational/polynomial formulation of the problem

3.1. Ubiquity of rational modeling

Let us remind that the signal reconstruction problem is tackled through the minimization of criterion $\mathcal{J}$ which has been defined in (5). We thus want to find

$$\mathcal{J}^* = \min_{\mathbf{x}\in\mathbb{R}^T} \quad \mathcal{J}(\mathbf{x})\,. \tag{7}$$

We emphasize that formulating our problem as a polynomial/rational one offers a widely applicable framework. First, let us show that there exists a polynomial reformulation of the ℓ_0 criterion. Indeed, choosing $\mathcal{R}_\lambda = \lambda\ell_0$ in the penalization

term of criterion (5), the original problem (7) can be reformulated by using a rational function and polynomial constraints, as follows:

$$\begin{aligned} \underset{(\mathbf{x},\boldsymbol{\xi})\in\mathbb{R}^T\times\mathbb{R}^T}{\text{minimize}} \quad & \|\mathbf{y} - D_\delta(\Phi(\mathbf{h} * (\mathbf{x} \odot \boldsymbol{\xi})))\|^2 + \lambda \sum_{t=1}^{T} \xi_t \\ \text{s.t.} \quad & (\forall t \in [\![1,T]\!]) \quad \xi_t = \xi_t^2 \,, \end{aligned} \tag{8}$$

where the operator $\odot$ denotes the element-wise Hadamard product. The ξ_i's are introduced to formulate the ℓ_0 penalization in a polynomial form, while the constraints ensure that they are binary variables. In this formulation unfortunately, both the number of variables and the degree of the involved polynomials are increased by a factor of two. As a consequence, we will show in Section 5 that (8) has a high complexity. The method presented in this article allows us to overcome this complexity barrier by using a different formulation of (5).

Looking closer at the relaxations of ℓ_0 mentioned in Section 2.2, an original alternative approach consists in considering penalization functions that are piecewise rational and can be expressed under the general form

$$(\forall x \in \mathbb{R}) \quad \Psi_\lambda(x) = \sum_{i=1}^{I} \zeta_i(x) \mathbb{1}_{\{\sigma_{i-1} \leq x < \sigma_i\}} \,, \tag{9}$$

where $(\zeta_i)_{i\in[\![1,I]\!]}$ are rational functions, I is a nonzero integer, and $(\sigma_i)_{i\in[\![0,I]\!]}$ is an increasing sequence of real values. The resulting criterion is thus a sum of rational and piecewise rational functions:

$$\begin{aligned} \mathcal{J}(\mathbf{x}) = & \sum_{u=1}^{U} g_u(x_{\Delta(u,\alpha)-L+1}, \ldots, x_{\Delta(u,\alpha)}) \\ & + \sum_{t=1}^{T} \sum_{i=1}^{I} \zeta_i(x_t) \mathbb{1}_{\{\sigma_{i-1} \leq x_t < \sigma_i\}} \,. \end{aligned} \tag{10}$$

3.2. Piecewise rational criteria

In this section, we first show how to transform the piecewise rational criterion in Problem (7) into the equivalent minimization of a sum of rational functions under polynomial constraints. To do so, we introduce the binary variables $\left(z^{(i)}\right)_{i\in[\![1,I]\!]}$ such that

$$(\forall i \in [\![0,I]\!]) \quad z^{(i)} = \mathbb{1}_{\{\sigma_i \leq x\}} \,.$$

We set $\sigma_0 = -\infty$, $z^{(0)} = 1$ and $\sigma_I = +\infty$, $z^{(I)} = 0$ to define Ψ_λ on the whole real line $\mathbb{R}$. From the definition of $\left(z^{(i)}\right)_{i\in[\![1,I]\!]}$, we deduce that

$$(\forall i \in [\![0,I]\!]) \quad \mathbb{1}_{\{\sigma_{i-1} \leq x < \sigma_i\}} = z^{(i-1)}(1 - z^{(i)}). \tag{11}$$

Finally, the constraint $z^{(i)} = \mathbb{1}_{\{\sigma_i \leq x\}}$ is equivalent to two polynomial constraints

$$z^{(i)} = \mathbb{1}_{\{\sigma_i \leq x\}} \Longleftrightarrow \begin{cases} \left(z^{(i)}\right)^2 - z^{(i)} = 0 \\ \left(z^{(i)} - \frac{1}{2}\right)(x - \sigma_i) \geq 0\,. \end{cases} \tag{12}$$

Indeed, the polynomial equality constraint enforces $z^{(i)}$ to be a binary variable while the polynomial inequality constraint ensures that it takes the same values as $\mathbb{1}_{\{\sigma_i \leq x\}}$ for every x in $\mathbb{R}$. Therefore, by substituting (12) for (10), Problem (7) reads as the minimization of a sum of rational functions depending on $\mathbf{x}$ and vectors $\mathbf{z} = \left(\mathbf{z}^{(i)}\right)_{i \in [\![0,I]\!]}$ under polynomial constraints, namely

$$\begin{aligned} \underset{\mathbf{x},\mathbf{z}}{\text{minimize}} \quad & \sum_{u=1}^{U} g_u(x_{\Delta(u,\alpha)-L+1}, \ldots, x_{\Delta(u,\alpha)}) \\ & + \sum_{t=1}^{T} \sum_{i=1}^{I} \zeta_i(x_t) z_t^{(i-1)} (1 - z_t^{(i)}) \\ \text{s.t.} \quad & (\forall (i,t) \in [\![0,I]\!] \times [\![1,T]\!]) \quad \begin{cases} \left(z_t^{(i)}\right)^2 - z_t^{(i)} = 0 \\ \left(z_t^{(i)} - \frac{1}{2}\right)(x_t - \sigma_{i+1}) \geq 0\,. \end{cases} \end{aligned} \tag{13}$$

More generally, this reformulation can be applied to the minimization of any piecewise rational function. For instance, a piecewise rational fit function $f_{\mathbf{y}}$ could also be chosen.

3.3. Symmetry of regularizers

All the piecewise rational approximations to the ℓ_0 penalty listed in Section 2.2 are even functions. This symmetry property is expressed here by an absolute value on the input variable x_t in the expressions of function Ψ_λ. This absolute value is handled in our framework by adding an additional variable r_t for each x_t and adding the two constraints

$$\begin{cases} r_t^2 = x_t^2 \\ r_t \geq 0\,. \end{cases} \tag{14}$$

This symmetry is important to decrease the number I of variables $\mathbf{z}$ involved in (13) and therefore to reduce the overall complexity of the final problem to be solved, as will be explained in Section 5. Indeed, it can divide by two the number of pieces in Ψ_λ, leading to only $I/2$ pieces instead of I. Taking the example of the MCP penalization, instead of having the four intervals, $]-\infty, -\gamma\lambda[$, $[-\gamma\lambda, 0[$, $[0, \gamma\lambda[$, and $[\gamma\lambda, +\infty[$, we have only the two intervals, $[0, \gamma\lambda[$ and $[\gamma\lambda, +\infty[$. Using symmetry results in adding one variable r_t and $I/2$ variables $\left(z_t^{(i)}\right)_{i \in [\![1,I/2]\!]}$ for each x_t as well as $2 + I/2$ polynomial constraints corresponding to constraints (12) and (14). This has to be compared with the

direct formulation where we introduce I variables $\left(z_t^{(i)}\right)_{i\in[\![1,I]\!]}$ with I polynomial constraints. Note that in our analysis of Section 5, we omit the equality constraints that force $\left(z_t^{(i)}\right)_{i\in[\![1,I/2]\!]}$ to be binary variables since substitution will be performed for those constraints in Section 5.2.

4. Solving the optimization problem

This section is concerned with the resolution of Problem (13) presented in a progressive manner. After a brief review of techniques from polynomial and rational optimization in Section 4.1, we apply the latter to our signal processing context. In Sections 4.2 and 4.3, we explicitly show how the structure of our problem allows us to reduce the dimensions of the final convex relaxation. Our analysis reveals that signal processing problems are computationally tractable when using sparsity patterns and subsampling.

4.1. Minimizing a rational function

A sum of rational functions can be written as a single rational function by reduction to a common denominator. A first step to handle (13) is hence to consider the minimization of a single rational function. In this section, we simplify our notation to explain the framework used to solve (13) and we focus on the generic problem of finding

$$\mathcal{J}^* = \min_{\mathbf{x}\in\mathcal{K}} \frac{p(\mathbf{x})}{q(\mathbf{x})}\,, \tag{15}$$

where p and q are polynomials in T variables and $\mathcal{K}\subset\mathbb{R}^T$ is the feasible set. Section 4.3 will get back to Problem (13).

4.1.1. Condition on the feasible set

In (15), $\mathcal{K}$ is a basic subset of $\mathbb{R}^T$ defined by polynomial inequalities as

$$\mathcal{K} = \{\mathbf{x}\in\mathbb{R}^T \mid (\forall j\in[\![1,J]\!]) \quad s_j(\mathbf{x})\geq 0\}\,, \tag{16}$$

where, for every $j\in[\![1,J]\!]$, $s_j\colon\mathbb{R}^T\to\mathbb{R}$. As we often work with bounded signals, we make the mild assumption that $\mathcal{K}$ contains T polynomial constraints of the form

$$(\forall t\in[\![1,T]\!]) \quad x_t^2\leq B^2\,,$$

where B is a positive constant. Since $\mathcal{K}$ is a closed set in a finite dimensional space, the above boundedness condition ensures that $\mathcal{K}$ is a compact set. To simplify the notation, we write those constraints into a vector form as

$$(\mathbf{B}-\mathbf{x})\odot(\mathbf{x}+\mathbf{B})\geq\mathbf{0}\,,$$

where $\mathbf{B}$ and $\mathbf{0}$ are the vectors composed solely of B and 0, respectively.

4.1.2. Reformulation as a moment problem

As shown in [34, Proposition 5.20], Problem (15) is equivalent to find

$$\begin{aligned} \inf_{\mu \in \mathcal{M}_+(\mathcal{K})} & \int_{\mathcal{K}} p(\mathbf{x}) \mu(\mathrm{d}\mathbf{x}) \\ \text{s.t.} & \int_{\mathcal{K}} q(\mathbf{x}) \mu(\mathrm{d}\mathbf{x}) = 1 \,, \end{aligned} \tag{17}$$

where $\mathcal{M}_+(\mathcal{K})$ denotes the set of positive finite measures supported on $\mathcal{K}$. The equivalence between Problems (15) and (17) relies on the possibility to link any optimal point $\mathbf{x}_*$ of (15) to a Dirac measure $\delta(\mathbf{x}_*)/q(\mathbf{x}_*)$ solution to (17). The main idea here is to embed the original problem in a higher dimensional space in order to linearize it. At first glance, (17) looks more intricate than (15) since we need to minimize over an infinite-dimensional set of measures supported by $\mathcal{K}$ instead of minimizing on $\mathcal{K}$ itself. However, the objective function and the constraint are now linear in the new optimized variable μ. Furthermore, by defining $\mathbf{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} \dots x_T^{\alpha_T}$, notice that

$$\int_{\mathcal{K}} p(\mathbf{x}) \mu(\mathrm{d}\mathbf{x}) = \int_{\mathcal{K}} \sum_{\boldsymbol{\alpha} \in \mathbb{N}^T} p_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}} \mu(\mathrm{d}\mathbf{x}) = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^T} p_{\boldsymbol{\alpha}} v_{\boldsymbol{\alpha}} \,, \tag{18}$$

where $v_{\boldsymbol{\alpha}} = \int_{\mathcal{K}} \mathbf{x}^{\boldsymbol{\alpha}} \mu(\mathrm{d}\mathbf{x})$ denotes the moment of order $\boldsymbol{\alpha}$ of the measure μ. For convenience, we will use infinite vectors to write sums such as the rightmost member of (18). We define the infinite vector $\widetilde{\mathbf{p}} = (p_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^T}$ and the infinite moment vector $\widetilde{\mathbf{v}} = (v_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^T}$. Since $\widetilde{\mathbf{p}}$ has a finite number of nonzero elements, the sum in (18) is well defined and can be written $\widetilde{\mathbf{p}}^\top \widetilde{\mathbf{v}}$.

Since $\mathcal{K}$ is a compact set, the measure μ is uniquely defined by its moments and thus we can reformulate Problem (17) as

$$\begin{aligned} \inf_{\widetilde{\mathbf{v}} \in \mathbb{R}^{\mathbb{N}^T}} \quad & \widetilde{\mathbf{p}}^\top \widetilde{\mathbf{v}} \\ \text{s.t.} \quad & \widetilde{\mathbf{q}}^\top \widetilde{\mathbf{v}} = 1 \\ & \widetilde{\mathbf{v}} \in \mathcal{D}(\mathcal{K}) \end{aligned} \,, \tag{19}$$

where $\widetilde{\mathbf{q}}$ is defined similarly to $\widetilde{\mathbf{p}}$ as the infinite vector extensions of $\mathbf{q}$ obtained by zero padding and $\mathcal{D}(\mathcal{K})$ is the cone of moments of positive measures supported on $\mathcal{K}$. Our objective now is to replace this difficult conic constraint by simpler constraints. We introduce two tools, respectively, the moment matrix $\mathbf{M}(\widetilde{\mathbf{v}})$ associated to the moment vector $\widetilde{\mathbf{v}}$ and the localizing matrix $\mathbf{M}^s(\widetilde{\mathbf{v}})$ associated to $\widetilde{\mathbf{v}}$ with respect to a given polynomial s. Those matrices are infinite-dimensional and are both defined through their entries as follows

$$\begin{aligned} &\left(\forall (\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathbb{N}^T \times \mathbb{N}^T\right) \quad M_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{v}) = v_{\boldsymbol{\alpha}+\boldsymbol{\beta}} \\ &\left(\forall (\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathbb{N}^T \times \mathbb{N}^T\right) \quad M^s_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{v}) = \sum_{\boldsymbol{\gamma} \in \mathbb{N}^T} s_{\boldsymbol{\gamma}} v_{\boldsymbol{\alpha}+\boldsymbol{\beta}+\boldsymbol{\gamma}} \,. \end{aligned}$$

We define such infinite-dimensional matrices to be positive semi-definite if all their finite-dimensional principal submatrices are positive semi-definite. Since $\mathcal{K}$ is compact, Putinar's theorem [35, Proposition 3.1] states that $\widetilde{\mathbf{v}}$ has a representing measure in $\mathcal{M}_+(\mathcal{K})$ if and only if the corresponding moment matrix $\mathbf{M}(\widetilde{\mathbf{v}})$ and localizing matrices $(\mathbf{M}^{s_j}(\widetilde{\mathbf{v}}))_{j\in[\![1,J]\!]}$ are positive semi-definite. The positive semi-definiteness of the moment and localizing matrices guarantee that $\widetilde{\mathbf{v}}$ represents a positive measure and ensures that its support is $\mathcal{K}$.

4.1.3. Converging hierarchy of SDP problems

To solve numerically Problem (19), we replace the conic constraint with semidefinite constraints and then truncate the moment vector $\widetilde{\mathbf{v}}$, as well as its associated moment and localizing matrices, up to a degree $2k$ for a given integer k. This yields a hierarchy of convex SDP problems, known as Lasserre's hierarchy [27]. For a given relaxation order k, the SDP relaxation to be solved reads:

$$\begin{aligned} \mathcal{J}_k^* = \quad & \inf_{\mathbf{v}\in\mathbb{R}^m} \quad \sum_{\boldsymbol{\alpha}\in\mathbb{N}_{2k}^T} p_{\boldsymbol{\alpha}} v_{\boldsymbol{\alpha}} \\ & \text{s.t.} \quad \sum_{\boldsymbol{\alpha}\in\mathbb{N}_{2k}^T} q_{\boldsymbol{\alpha}} v_{\boldsymbol{\alpha}} = 1 \\ & \qquad \mathbf{M}_k(\mathbf{v}) \in \mathbb{S}_+^{n_0} \\ & \qquad (\forall j \in [\![1,J]\!]) \quad \mathbf{M}_{k-\mathrm{d}_{s_j}}^{s_j}(\mathbf{v}) \in \mathbb{S}_+^{n_j}\,, \end{aligned} \tag{20}$$

where $(\mathrm{d}_{s_j})_{j\in[\![1,J]\!]}$ are defined in (1). The cardinality of $\mathbb{N}_{2k}^T$ is $\binom{T+2k}{2k}$ and thus $\mathbf{v}$ is a vector containing $m = \binom{T+2k}{2k}$ moments. Furthermore, the truncated moment matrix $\mathbf{M}_k$ is the principal submatrix of the infinite-dimensional moment matrix $\mathbf{M}$ that has dimension $n_0 \times n_0$ with $n_0 = \binom{T+k}{k}$. Thereby, $\mathbf{M}_k$ is indexed by $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in $\mathbb{N}_k^T \times \mathbb{N}_k^T$ and contains all the moments up to degree $2k$. Similarly, the truncated localizing matrices are the submatrix of their infinite-dimensional counterparts that have the dimensions $n_j \times n_j$ with $n_j = \binom{T+k-\mathrm{d}_{s_j}}{k-\mathrm{d}_{s_j}}$.

Problem (20) is an SDP problem in its dual form with linear equality constraints. Indeed, aggregating the moment and localizing matrices into a single symmetric block diagonal matrix before separating it into a sum along the elements of $\mathbf{v}$, we obtain

$$\begin{aligned} \mathcal{J}_k^* = \quad & \underset{\mathbf{v}\in\mathbb{R}^m}{\text{minimize}} \quad \mathbf{p}^\top \mathbf{v} \\ & \text{s.t.} \quad \mathbf{C} - \sum_{i=1}^m v_i \mathbf{A}_i \in \mathbb{S}_+^n \\ & \qquad \mathbf{a} - \mathbf{G}^\top \mathbf{v} = \mathbf{0}\,, \end{aligned} \tag{21}$$

where $\mathbf{C}$ and $(\mathbf{A}_i)_{i\in[\![1,m]\!]}$ are symmetric matrices, $\mathbf{a}$ is a vector of $\mathbb{R}^\ell$, and $\mathbf{G}$ is a matrix of $\mathbb{R}^{m\times\ell}$. The dimension n is thus given by $n = \sum_{j=0}^J n_j$. Notice that, in this section, (20) has only one linear constraint thus $\ell = 1$, $\mathbf{a} = 1$, and

$\mathbf{G} = \mathbf{q}$. However, in next sections, more linear constraints will be involved, so that we prefer to employ this matrix-vector notation here.

Solving each SDP problem yields a lower bound $\mathcal{J}_k^*$ on the optimal value $\mathcal{J}^*$ of the criterion $\mathcal{J}$. Furthermore, the higher the order k, the tighter the bound $\mathcal{J}_k^*$ but the higher also the dimensions of the SDP problem. In our context where the sought signal is bounded, $(\mathcal{J}_k^*)_{k\in\mathbb{N}}$ is an increasing convergent sequence whose limit is $\mathcal{J}^*$ [27]. Moreover, the hierarchy has finite convergence generically, i.e. convergence happens at a finite relaxation order generically [36]. Finally, an exact global solution $\hat{\mathbf{x}}$ of (15) can be extracted from the solution of an SDP problem (20) indexed by a relaxation order at which convergence has occurred [37].

Note that the relaxation order k should be chosen such that

$$k \geq \max\left\{\mathrm{d}_p, \mathrm{d}_q, \max_{j\in[\![1,J]\!]} \mathrm{d}_{s_j}\right\} .$$

This is a necessary condition which ensures that $2k$ is greater than the maximum degree of p, q and all the $(s_j)_{j\in[\![1,J]\!]}$, and prevents truncation of the latter polynomials. There is no a priori known sufficient relaxation order to ensure the convergence of the hierarchy. However, once the SDP relaxation is solved, there exists a sufficient condition that guarantees the convergence. Namely, if the moment matrices $\mathbf{M}_k$ and $\mathbf{M}_{k-1}$ have the same rank, then convergence has occurred [34].

We remark that the dimensions n and m of the SDP problem grows respectively as T^k and T^{2k} when T is large, hence exponentially in the degree of the involved polynomials.

4.2. Problem structure emerging from a sum of rational functions

Although constituting the theoretical foundation of our work, the approach presented in Section 4.1 is computationally inefficient and requires further improvements that we now explain. Indeed, reducing a sum of rational functions using a common denominator often yields a rational function with very high degree, which then requires a high relaxation order k in the hierarchy of SDP problems. As a consequence, the obtained SDP problems are too high-dimensional to be solvable in a reasonable time using state-of-the-art solvers. However, a more ingenious method is to use the structure induced by the sum to yield a block SDP problem [28]. There are two types of structure to consider in our problem: first we deal with a sum of rational functions instead of a single one, and then each of those functions has only a few subset of variables as input. Those two structures are sometimes referred to as sparse problem and sparse polynomials [38, 39]. However, in order to prevent confusion with the sparsity of the original signal $\overline{\mathbf{x}}$, we will not use this terminology. To illustrate it, let us turn our attention on finding

$$\mathcal{J}^* = \min_{\mathbf{x}\in\mathcal{K}} \sum_{i=1}^{\tilde{I}} \frac{p_i(\mathbf{x}_{E_i})}{q_i(\mathbf{x}_{E_i})} , \tag{22}$$

where p_i and q_i are polynomials in T_i variables and $\mathcal{K}$ is a compact subset of $\mathbb{R}^T$ having the form (16). The vector $\mathbf{x}_{E_i}$ denotes the subvector of $\mathbf{x}$ composed of the elements indexed by the set E_i, the set E_i being a subset of $[\![1, T]\!]$ of cardinality T_i. We further assume that the polynomials in (22) involve only a few variables, i.e.

$$\left(\forall i \in [\![1, \tilde{I}]\!]\right) \quad T_i \ll T\,. \tag{23}$$

4.2.1. Exploiting the sum of rational functions structure

Instead of introducing a single measure on all the variables, we now introduce a measure μ_i for each rational function p_i/q_i of the sum. However, coupling between variables from different measures appears when the sets $(E_i)_{i\in[\![1,\tilde{I}]\!]}$ intersect. We therefore need to add moment equality constraints to ensure that the overlapping moments of two measures are identical. Moreover, we need some restrictions on how variables can overlap several measures in order to keep the problem consistent. The required condition is that the sets $(E_i)_{i\in[\![1,\tilde{I}]\!]}$ verify the so-called running intersection property [34, 39] which is stated as

$$(\forall i \in [\![2, \tilde{I}]\!], \exists \tilde{\jmath} \in [\![1, i]\!]) \quad E_i \bigcap \left(\bigcup_{j=1}^{i-1} E_j \right) \subseteq E_{\tilde{\jmath}}\,.$$

Together with the compactness of $\mathcal{K}$, it guarantees that Problem (22) is equivalent to

$$\begin{aligned}
&\inf_{\boldsymbol{\mu}\in\Xi} \sum_{i=1}^{\tilde{I}} \int_{\mathcal{K}_i} p_i(\mathbf{x}_{E_i})\mu_i(\mathrm{d}\mathbf{x}_{E_i}) \\
&\text{s.t. } (\forall i \in [\![1, \tilde{I}]\!]) \int_{\mathcal{K}_i} q_i(\mathbf{x}_{E_i})\mu_i(\mathrm{d}\mathbf{x}_{E_i}) = 1 \\
&\qquad (\forall (i,j) \in [\![1, \tilde{I}]\!] \times [\![1, \tilde{I}]\!])(\forall \boldsymbol{\gamma} \in \mathbb{N}^{\mathrm{Card}(E_{i,j})}) \\
&\qquad \int_{\mathcal{K}_i} q_i(\mathbf{x}_{E_i})\mathbf{x}_{E_{i,j}}^{\boldsymbol{\gamma}} \mu_i(\mathrm{d}\mathbf{x}_{E_i}) = \int_{\mathcal{K}_j} q_j(\mathbf{x}_{E_j})\mathbf{x}_{E_{i,j}}^{\boldsymbol{\gamma}} \mu_j(\mathrm{d}\mathbf{x}_{E_j})\,,
\end{aligned} \tag{24}$$

where $E_{i,j} = E_i \bigcap E_j$ and $\boldsymbol{\mu} = (\mu_i)_{i\in[\![1,\tilde{I}]\!]}$ is the new optimization variable belonging to the product $\Xi = \bigtimes_{i\in[\![1,\tilde{I}]\!]} \mathcal{M}_+(\mathcal{K}_i)$. The sets $(\mathcal{K}_i)_{i\in[\![1,\tilde{I}]\!]}$ are subsets of $\mathbb{R}^{T_i}$ defined by the subsets of polynomials in variables $\mathbf{x}_{E_i}$ defining $\mathcal{K}$. The last equality constraints in (24) enforce equality between the marginal distributions of $q_j\mu_j$ and $q_i\mu_i$ along $\mathbf{x}_{E_{ij}}$. In other words, those constraints ensure the equality of overlapping moments between the different measures.

4.2.2. Block structure in the SDP hierarchy

As in Section 4.1, we now use Putinar's theorem to replace each measure by its moment vector at the cost of additional semi-definite constraints. We then truncate the moment vectors as well as the moment and localizing matrices, before stacking them. As a result, the moment vector $\mathbf{v} = [\mathbf{v}_1^\top, \ldots, \mathbf{v}_{\tilde{I}}^\top]^\top$ is

a stack of the moment vectors of each measure μ_i. Similarly, the moment matrix $\mathbf{M}_k(\mathbf{v}) = \mathrm{Diag}\left(\mathbf{M}_{1,k}(\mathbf{v}_1), \ldots, \mathbf{M}_{\tilde{I},k}(\mathbf{v}_{\tilde{I}})\right)$ and the localizing matrices $\mathbf{M}^{s_j}_{k-\mathrm{d}_{s_j}}(\mathbf{v}) = \mathrm{Diag}\left(\mathbf{M}^{s_j}_{1,k-\mathrm{d}_{s_j}}(\mathbf{v}_1), \ldots, \mathbf{M}^{s_j}_{\tilde{I},k-\mathrm{d}_{s_j}}(\mathbf{v}_{\tilde{I}})\right)$ have a block diagonal structure where each diagonal block corresponds respectively to the moment or localizing matrix of one of the measures μ_i. This leads to the following SDP problem:

$$\begin{aligned}
\mathcal{J}_k^* = \inf_{\mathbf{v}\in\mathbb{R}^m} \; & \mathbf{p}^\top \mathbf{v} \\
\text{s.t. } (\forall i \in [\![1, \tilde{I}]\!]) \quad & \mathbf{q}_i{}^\top \mathbf{v}_i = 1 \\
& \mathbf{M}_k(\mathbf{v}) \in \mathbb{S}_+^{n_0} \\
& (\forall j \in [\![1, J]\!]) \quad \mathbf{M}^{s_j}_{k-\mathrm{d}_{s_i}}(\mathbf{v}) \in \mathbb{S}_+^{n_j} \\
& \mathbf{F}\mathbf{v} = \mathbf{0}\,,
\end{aligned} \tag{25}$$

where

$$m = \sum_{i=1}^{I} \binom{T_i + 2k}{2k}, \; n_0 = \sum_{i=1}^{I} \binom{T_i + k}{k}, \; n_j = \sum_{i=1}^{I} \binom{T_i + k - \mathrm{d}_{s_j}}{k - \mathrm{d}_{s_j}},$$
$$\mathbf{p} = [\mathbf{p}_1^\top, \ldots, \mathbf{p}_{\tilde{I}}^\top]^\top,$$

and $\mathbf{F}$ is a matrix in $\mathbb{R}^{\ell\times m}$ representing the linear constraints linking the $(\mathbf{v}_i)_{i\in[\![1,\tilde{I}]\!]}$ together and coming from the constraints between the projections in (24). Similarly to Section 4.1, (25) can be finally expressed in the canonical form (21).

There are two main differences with the situation discussed in Section 4.1:

- Instead of having a single measure on all the variables, we obtain several measures on different smaller subsets of variables. The SDP optimization variable $\mathbf{v}$ is now a vector built by stacking the different truncated moment vectors $(\mathbf{v}_i)_{i\in[\![1,\tilde{I}]\!]}$ of each measure. As a consequence, the moment and localizing matrices have a block diagonal structure, each block corresponding to a measure, or equivalently to a term in the sum of Problem (22). Thanks to Assumption (23), the size of the blocks in the moment matrix, equal to $\binom{T_i+2k}{2k}$, is much smaller than the size $\binom{T+2k}{2k}$ obtained in Section 4.1. Especially when k increases, the difference in size becomes even more significant. The block structure can then be efficiently exploited by SDP solvers to decrease the computational time.

- Extra moment constraints due to the coupling between variables arise. Although those constraints may be numerous, they are linear equality constraints in the SDP problem; their impact on the computational time of the SDP solver is minor.

4.3. Minimizing our criterion

We now apply the method of Section 4.2 to solve (13). We have to handle a sum of $\tilde{I} = U + T$ terms. We hence introduce a measure for each term, i.e.

U measures $(\mu_u)_{u\in[\![1,U]\!]}$ for the rational functions $(g_u)_{u\in[\![1,U]\!]}$ and T measures $(\nu_t)_{t\in[\![1,T]\!]}$ for the rational functions in the reformulated penalization. The measures $(\mu_u)_{u\in[\![1,U]\!]}$ are measures on at most L variables $x_{\Delta(u,\alpha)-L+1},\dots,x_{\Delta(u,\alpha)}$ while the measures $(\nu_t)_{t\in[\![1,T]\!]}$ are measures on $I+1$ scalar variables, corresponding to x_t and $\mathbf{z}_t$.

4.3.1. Feasible set for our reformulated problem

In Problem (13), the sets $(\mathcal{K}_i)_{i\in[\![1,U+T]\!]}$ are defined by the bound constraints and the polynomial constraints arising from the reformulation of Section 3.2. Namely, the sets $(\mathcal{K}_i)_{i\in[\![1,U]\!]}$ are defined by

$$(\forall i\in[\![1,U]\!])\quad (\mathbf{B}-\mathbf{x}_{E_i})\odot(\mathbf{x}_{E_i}+\mathbf{B})\geq \mathbf{0}\,, \tag{26}$$

while the sets $(\mathcal{K}_i)_{i\in[\![U+1,U+T]\!]}$ are defined by

$$\begin{aligned}(\forall i\in[\![U+1,U+T]\!])\quad & (B-x_{i-U})(x_{i-U}+B)\geq 0\\ & (\forall j\in[\![1,I]\!])\quad \left(z_{i-U}^{(j)}\right)^2-z_{i-U}^{(j)}=0\\ & (\forall j\in[\![1,I]\!])\quad \left(z_{i-U}^{(j)}-\frac{1}{2}\right)(x_{i-U}-\sigma_{j+1})\geq 0\,.\end{aligned} \tag{27}$$

Note that, since we introduce a measure for each rational function in the sum, we have to cope with more than T bound constraints. Indeed several measures are defined on identical variables and we need to introduce bound constraints for each of those measures. We then perform the relaxation (25) to generate a hierarchy of SDP problems.

4.3.2. Coupling and linear equality constraints

We observe that two kinds of coupling as discussed in Section 4.2.1 appear: one between the different measures $(\mu_u)_{u\in[\![1,U]\!]}$ and one between the measures $(\mu_u)_{u\in[\![1,U]\!]}$ and the measures $(\nu_t)_{t\in[\![1,T]\!]}$. By definition of the convolution matrix $\mathbf{H}$ in Section 2, the sets $(E_i)_{i\in[\![1,\tilde{I}]\!]}$ satisfy the running intersection property. Furthermore, among the extra moment equality constraints to acknowledge coupling, we remark that many of them between moments of the measures $(\mu_u)_{u\in[\![1,U]\!]}$ are redundant. Let us take a simple example to illustrate this fact.

Example. Assume that we want to minimize over the variable $\mathbf{x}=(x_t)_{t\in[\![1,5]\!]}$, a sum of three rational functions which has the following form

$$\frac{p_1(x_1,x_2,x_3)}{q_1(x_1,x_2,x_3)}+\frac{p_2(x_2,x_3,x_4)}{q_2(x_2,x_3,x_4)}+\frac{p_3(x_3,x_4,x_5)}{q_3(x_3,x_4,x_5)}\,.$$

Following the method developed in Section 4.2, we introduce three measures μ_1 μ_2 and μ_3, one for each term of the sum. We thus need the following equality

constraints between moments, for every (α, β) in $\mathbb{N}^2$,

$$\int q_1(x_1, x_2, x_3) x_2^\alpha x_3^\beta \mu_1(\mathrm{d}x_1, \mathrm{d}x_2, \mathrm{d}x_3) = \int q_2(x_2, x_3, x_4) x_2^\alpha x_3^\beta \mu_2(\mathrm{d}x_2, \mathrm{d}x_3, \mathrm{d}x_4)$$
$$\int q_2(x_2, x_3, x_4) x_3^\alpha x_4^\beta \mu_2(\mathrm{d}x_2, \mathrm{d}x_3, \mathrm{d}x_4) = \int q_3(x_3, x_4, x_5) x_3^\alpha x_4^\beta \mu_3(\mathrm{d}x_3, \mathrm{d}x_4, \mathrm{d}x_5)$$
$$\int q_1(x_1, x_2, x_3) x_3^\alpha \mu_1(\mathrm{d}x_1, \mathrm{d}x_2, \mathrm{d}x_3) = \int q_3(x_3, x_4, x_5) x_3^\alpha \mu_3(\mathrm{d}x_3, \mathrm{d}x_4, \mathrm{d}x_5)\,.$$

We observe that the variable x_3 appears in each term of the sum and thus also in moments of each measure. In particular, we notice that the last constraint is redundant with the first two ones when $\beta = 0$. It is thus sufficient to consider only moment equality constraints on consecutive measures $(\mu_u)_{u \in [\![1,U]\!]}$ in (13). We can thus drastically reduce the number of moment linear equality constraints.

5. Complexity of the relaxation

Current state-of-the-art SDP solvers use interior points methods which are known to be very efficient for small and medium scale problems. On the other hand, their running time becomes prohibitive for large scale problems. This is a major drawback of the relaxation of polynomial optimization problems into SDP problems. Nevertheless, Sections 4.2 and 4.3 used the structure of the problem to yield a structured SDP problem. In this section, we derive the complexity of this SDP problem and show that it is computationally solvable in a fair amount of time.

The complexity of an SDP problem under the form (21) is expressed as a quadruple of integers $(n, m, m_\mathrm{s}, \ell)$. The integer m denotes the size of the vector of optimized variables, n is the size of the semi-definite inequality constraint, ℓ is the number of linear equality constraints, and m_s is the number of block matrices involved in the semi-definite constraint. Note that n is related to m_s since it is the sum of the size of each block. The above quadruple therefore does not fully characterize the structure of an SDP. For example, having one huge block and nine tiny ones is not equivalent in terms of complexity to having ten medium blocks. However, knowing n and m_s is usually enough to get a good evaluation of the complexity of the problem.

The bottleneck for current SDP solver is mainly the dimension of both n and m. This section gives an asymptotic estimation for n and m according to the parameters of our initial model (2) and the relaxation order k. A more detailed derivation for the expression of $(n, m, m_\mathrm{s}, \ell)$ is presented in Appendix A. We show first that subsampling and sparsity allow to decrease the latter and make the numerical resolution of the associated SDP problems tractable. Then, we introduce tools and tricks that allow us to decrease further the dimension of the SDP problems to be solved, so reducing the computational time of our method.

5.1. Consequence of the subsampling on the dimensions of the SDP problem

For a given relaxation order k, when the number of samples T goes to infinity and $L \gg 1$ (i.e. we lose the band structure of $\mathbf{H}$), the size of the SDP problem

asymptotically becomes of the order (see Appendix A)

$$m = \mathcal{O}(UL^{2k} + T) \quad , \quad n = \mathcal{O}(UL^k + T)\,. \tag{28}$$

We note that both sizes n and m grow exponentially with k and blow up quickly. In particular, m grows faster than n. However, we will see that the SDP hierarchy often converges quickly in practice, that is $(\mathcal{J}_k^*)_{k\in\mathbb{N}}$ converge to $\mathcal{J}^*$ for a relaxation order k of 2, 3, or 4. From our analysis, we observe that the main bottleneck of our method is the number of variables per measure and the order of relaxation. While the number of variables in measures $(\nu_t)_{t\in[\![1,T]\!]}$ is fixed to $I+1$, the total number of variables in measures $(\mu_u)_{u\in[\![1,U]\!]}$ is L and (28) shows that m and n rise quickly with L.

Although the subsampling reduces the quality of the reconstruction by eliminating some information on the signal, as a side effect in our context, it allows the size of the SDP relaxation to be reduced. As shown by (4), decimation decreases U, which plays a prominent role in the complexity parameters $(n, m, m_{\mathrm{s}}, \ell)$ of the SDP problem. Table 1 compares the size of SDP relaxations for the SCAD penalization without decimation (D_∞) and with D_2 (resp. D_4) decimation. As discussed above, the dimensions n and m increase quickly with the relaxation order k and the length of the filter L. Note that because of the approximation made in Section A.1, stating that measures $(\mu_u)_{u\in[\![1,U]\!]}$ are on L variables, the SDP dimension presented here are slightly overestimated.

Table 1: Dimension of the relaxation of the SCAD penalization for different decimations

			m			n			m_{s}			ℓ		
T	L	k	D_∞	D_4	D_2	D_∞	D_4	D_2	D_∞	D_4	D_2	D_∞	D_4	D_2
50	3	3	8400	7476	6300	7000	6450	5750	600	556	500	1035	735	420
100	3	3	16800	14784	12600	14000	12800	11500	1200	1104	1000	2085	1755	845
50	4	3	14700	12390	9450	9250	8205	6875	650	595	425	2015	1355	660
100	4	3	29400	24360	18900	18500	16220	13750	1300	1180	1050	4065	3405	1335
100	5	3	54600	43512	31500	25600	21236	17050	1400	1256	1100	7530	6375	2315
50	3	4	16500	14685	12375	13500	12455	11125	600	556	500	1772	1184	568
100	3	4	33000	29040	24750	27000	24720	22250	1200	1104	1000	3572	2102	1143
100	4	4	66000	54120	41250	38500	33460	28000	1300	1180	1050	9116	7268	2172

5.2. Polynomial equality constraints and substitution

For a given measure, equality constraints involving monic monomials in the definition of the support set $\mathcal{K}_i$ can be substituted. The constraint is then used to reduce the number of moments in the vector of moments. We clarify this process here through the example of the SCAD penalization. Substitution is carried out automatically by some software [40], but has not been clearly documented.

Let us focus our attention on the measure ν_t, depending on the three variables x_t, $z_t^{(1)}$, and $z_t^{(2)}$, as well as on the associated truncated vector $\mathbf{v}_t$ of moment up to degree 2. Using the equality constraints in (27), we substitute the related monomial in $\mathbf{v}_t$. The moments associated with monomials $\left(z_t^{(1)}\right)^2$ and $\left(z_t^{(2)}\right)^2$ are thus the same as the ones associated with $z_t^{(1)}$ and $z_t^{(2)}$. Therefore, the moment vector $\mathbf{v}_t$ has a dimension reduced by two. When $\mathbf{v}_t$ contains moments up to degree $2k$, substitution reduces the number of moments from $\binom{3+2k}{2k}$ to $8k$.

In the general case, for a given relaxation order k, $\mathbf{v}_t$ contains only $2k(k+1)$ moments after substitution which is much fewer than the original $\binom{1+I+2k}{2k}$ moments. Substitution significantly decreases the values of n, m and m_{s} which have a major impact on the computational cost of SDP solvers. However, it does not impact the number of linear constraints ℓ.

5.3. Linear versus quadratic polynomial constraints

Our method to solve rational optimization problem is valid only if the constraint set $\mathcal{K}$ is compact. We therefore set a bound B on the sought signal. The bound constraints can be expressed in two ways:

- first, as two linear vector constraints

$$\begin{aligned}\mathbf{B}-\mathbf{x} &\geq \mathbf{0}\\ \mathbf{x}+\mathbf{B} &\geq \mathbf{0}\,,\end{aligned}$$

- or as a single quadratic vector constraint

$$(\mathbf{B}-\mathbf{x})\odot(\mathbf{x}+\mathbf{B}) \geq \mathbf{0}\,.$$

Following Appendix A.1, using two linear inequality constraints per variable introduces $2(U+L)$ localizing matrices and consequently $2(U+L)$ blocks in our SDP problems while using a quadratic inequality constraint only adds $U+L$ blocks. Moreover, linear and quadratic constraints yield blocks of identical size. Indeed, the size of a localizing matrix $\mathbf{M}_k^s$ corresponding to a polynomial s in ω variables is given by $\binom{\omega+k-\mathrm{d}_s}{k-\mathrm{d}_s}$ and here $\mathrm{d}_s=1$ for both linear and quadratic constraints. Therefore, formulating the bound constraints as quadratic constraints reduces by a factor two the number of blocks associated to such bounds.

5.4. Using a sign oracle

For real-valued signals $\overline{\mathbf{x}}$, convergence is observed at orders k for which building and solving the corresponding SDP problems is highly demanding in terms of computation and memory storage. Conversely, when $\overline{\mathbf{x}}$ is a positive signal, we observed [28] convergence at a lower order k. This suggests a method yielding similar results for real-valued signals using an oracle. Instead of

$$(\forall \mathbf{x}\in\mathbb{R}^T)\quad \mathcal{J}(\mathbf{x}) = \frac{1}{2}\left\|\mathbf{y}-D_\delta(\Phi(\mathbf{H}\mathbf{x}))\right\|^2 + \sum_{t=1}^{T}\Psi_\lambda(x_t)\,,$$

we minimize

$$(\forall \mathbf{x} \in \mathbb{R}_+^T) \quad \widetilde{\mathcal{J}}(\mathbf{x}) = \frac{1}{2} \left\| \mathbf{y} - D_\delta(\Phi(\widetilde{\mathbf{H}}\mathbf{x})) \right\|^2 + \sum_{t=1}^{T} \Psi_\lambda(x_t) \,,$$

where $\widetilde{\mathbf{H}} = \mathbf{H}\,\mathrm{Diag}(\boldsymbol{\epsilon})$, $\boldsymbol{\epsilon} \in \{-1, 1\}^T$ is the sign vector of $\overline{\mathbf{x}}$ provided by the oracle, and $\mathrm{Diag}(\boldsymbol{\epsilon})$ is a diagonal matrix with binary elements $\boldsymbol{\epsilon}$. We build our oracle by solving a standard least absolute shrinkage and selection operator (LASSO) problem [6], i.e. $\Psi_\lambda = \lambda|.|$. The availability of an oracle allows us to restrict the minimization of (10) to positive valued signals thanks to the new convolution matrix $\widetilde{\mathbf{H}}$. Our oracle decreases significantly the computational time in two ways:

- Since the convergence of the SDP hierarchy occurs for smaller order k, the dimensions of the SDP problem to solve are much lower according to Section 5.
- Moreover, since we optimize now on positive variables, we do not need to use the additional variables $(r_t)_{t \in [\![1,T]\!]}$ introduced in Section 3.3 to account for symmetries and the presence of absolute values. This results in smaller vectors of moments, hence a lower dimensional SDP problem.

An exact solution is thus retrieved by solving an SDP problem of fair dimension. Finally, the computational cost of our oracle is low since we solve a LASSO using a forward-backward algorithm. It typically takes less than a second which is negligible compared to the computational time of our method as shown in Section 6 while providing accurate oracle on the sign of the initial signal.

6. Numerical simulations and results

6.1. Experimental set-up

To show the efficiency of our framework, we apply it to the reconstruction of a sparse signal subject to nonlinear distortion and subsampling. We use a piecewise relaxation of ℓ_0 to promote sparsity as detailed in Section 2.2. We perform simulations on 50 test cases where the initial sparse signal $\overline{\mathbf{x}}$ has length $T = 100$ with 10 non-zero values. Those values are drawn randomly according to a uniform distribution on $[-1, -0.1] \cup [0.1, 1]$. The position of the non-zero values are also drawn randomly according to uniform distribution on $[\![1, T]\!]$. The length L of the filter is set to 3 and its coefficients are the normalized L-th row of Pascal's triangle. This kind of filter is useful to model enlargement due to measurement from sensors for example. We choose the following saturation function for the nonlinear distortion ϕ

$$(\forall t \in \mathbb{R}) \quad \phi(t) = \frac{t}{\chi + |t|} \,,$$

where χ is set to 0.3. Finally, we perform the relaxation into SDP for relaxation orders 2, 3, and 4. We use GloptiPoly [40] to relax rational problems into SDP

problems which are then solved with the solver SDPT3 [41]. All the simulations have been run on a standard computer with an Intel Xeon CPU running at 3.7 GHz and 32 GB of RAM allocated to the process.

6.2. Example of a rational relaxation: SCAD

To clarify the reformulation of Section 3.2, we demonstrate it on the regularizers given in Section 2.2. Taking advantage of symmetry as explained in Section 3.3, SCAD has three pieces and thus requires to introduce variables $z_t^{(1)}$ and $z_t^{(2)}$ leading to

$$\begin{aligned}
\underset{\mathbf{x},\mathbf{z}}{\text{minimize}} \quad & f_{\mathbf{y}}(\mathbf{x}) + \sum_{t=1}^{T} (1 - z_t^{(1)})\lambda\,|x_t| + z_t^{(2)} \frac{(\gamma+1)\lambda^2}{2} \\
& - z_t^{(1)}(1 - z_t^{(2)}) \frac{\lambda^2 - 2\gamma\lambda\,|x_t| + x_t^2}{2(\gamma-1)} \\
\text{s.t.} \quad & (\forall (i,t) \in \{1,2\} \times [\![1,T]\!]) \quad \left(z_t^{(i)}\right)^2 - z_t^{(i)} = 0 \\
& (\forall t \in [\![1,T]\!]) \quad \left(z_t^{(1)} - \frac{1}{2}\right)(|x_t| - \gamma\lambda) \geq 0 \\
& (\forall t \in [\![1,T]\!]) \quad \left(z_t^{(2)} - \frac{1}{2}\right)(|x_t| - \lambda) \geq 0\,.
\end{aligned} \tag{29}$$

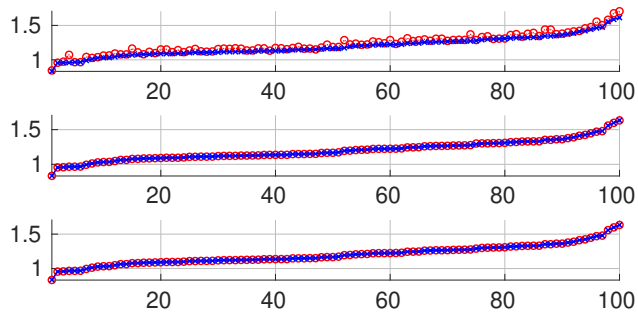
A similar approach applies to Capped ℓ_p, MCP, and CEL0 penalties; the details are omitted for conciseness. Although we use SCAD penalization in all the subsequent simulations, similar results can be obtained with Capped ℓ_p, MCP, and CEL0. Nonetheless, SCAD is more demanding in terms of computation since it has more rational pieces. It consequently provides a worst case scenario for the computational time compared with the other penalizations. The parameter γ for SCAD is set to 2.1 in order to approximate ℓ_0 closely. The value of the parameter λ was determined empirically and set to 0.15.

6.3. Acceleration of convergence with the sign oracle

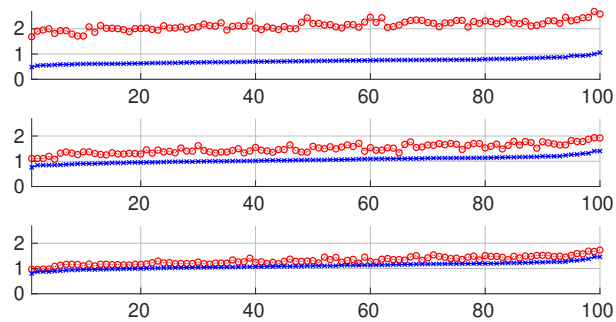
In this section, we want to show how the oracle impacts the convergence of the SDP hierarchy. We first consider the use of a sign oracle in a linear model, i.e. the case when $\phi = \mathrm{Id}$. We then delve into the more challenging case of a nonlinear model. Decimation is set to D_4 in this section. The oracle is build on solving a LASSO problem by using a forward-backward algorithm as described in Section 5.4.

6.3.1. Linear case

Solving each SDP problem in the hierarchy provides both a lower bound $\mathcal{J}_k^*$, which is the value of the objective function of the SDP at optimality, and an approximate minimizer $\hat{\mathbf{x}}_k$, which is extracted from a minimizer of the SDP problem. We compare here the value of the criterion at $\hat{\mathbf{x}}_k$ with $\mathcal{J}_k^*$. Since increasing the relaxation order k yields larger lower bounds and smaller criterion



(a) With oracle: $k = 2$ (top), 3 (middle) 4 (bottom)



(b) Without oracle: $k = 2$ (top), 3 (middle) 4 (bottom)

Figure 2: Comparison between the lower bound $\mathcal{J}_k^*$ and the value of the criterion $\mathcal{J}(\hat{\mathbf{x}}_k)$ for 100 tests (Linear Case).

values, we know that the convergence of the hierarchy happens when $\mathcal{J}(\hat{\mathbf{x}}_k)$ and $\mathcal{J}_k^*$ are equal. Figure 2 compares those two values respectively in the cases with oracle and without the use of our oracle on 100 test cases. From top to bottom, the two figures are drawn for relaxation orders $k = 2$, $k = 3$, and $k = 4$. Criterion values are represented in red while lower bounds are represented in blue. Each point of the x-axis represents the values for a single test case. For the sake of clarity, the values are ordered according to the value of the lower bound. We observe that, without oracle, the convergence is slow and still not reached in general at order $k = 4$. On the other hand, when we use our oracle, convergence appears quickly, i.e. $k = 3$ in most of the test cases.

6.3.2. Nonlinear case

Figure 3 is similar to Figure 2 but in the context of a nonlinear model. The continuous line with cross dots represents the cases without the use of an oracle while the dashed line with circle dots represents the cases with our sign oracle.

We observe here that even with a sign oracle, the convergence of the hierarchy does not occur for low values of k due to the nonlinearity. However, we can notice that the gap between the lower bound and the criterion value at the $\hat{\mathbf{x}}_k$ is greatly reduced when we use our oracle.

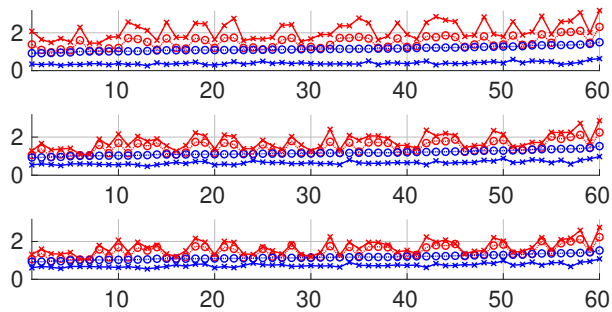


Figure 3: Comparison between the lower bound $\mathcal{J}_k^*$ and the value of the criterion $\mathcal{J}(\hat{\mathbf{x}}_k)$ when the oracle is used. Plain line with cross dots: no oracle used, dashed line with circle dots: use of our oracle (Nonlinear case).

6.4. Reconstruction of sparse signals

6.4.1. Global optimality

In this section, we want to demonstrate the quality of the minimizers of Problem (29) returned by various methods. Note that we do not use the oracle here. We use the decimation operator D_4 but similar results hold for the other operators. We compare our method to a forward-backward (FB) algorithm applied directly to the criterion $\mathcal{J} = f_{\mathbf{y}} + \mathcal{R}_\lambda$ where the gradient step is first performed on the data fitting term and a proximal step is then performed on the penalization. Hence the criterion to minimize is the same for both methods. We initialize the FB algorithm first with the null vector and denote by $\mathbf{x}_{\text{FB0}}$ the resulting solution. Then we perform a warm start of the FB algorithm using the solution obtained from our method as an initializer. The resulting estimate is denoted by $\mathbf{x}_{\text{FB1}}$.

In Figure 4, we compare the value of the criterion $\mathcal{J}$ at $\mathbf{x}_{\text{FB0}}$ and $\mathbf{x}_{\text{FB1}}$ with the solution returned by our method for a relaxation order $k = 4$. The solid blue curve with cross dots represents the values of the lower bound $\mathcal{J}_4^*$, the pointed red curve with circle dots represents $\mathcal{J}(\hat{\mathbf{x}}_4)$, the dashed green curve with plus dots represents $\mathcal{J}(\mathbf{x}_{\text{FB0}})$, and the dashed purple curve with plus dots represents $\mathcal{J}(\mathbf{x}_{\text{FB1}})$.

Since the criterion $\mathcal{J}$ is highly nonconvex, the forward-backward algorithm gets stuck in local minimizers. Indeed, changing the initialization point changes the output of the algorithm. We can observe it on Figure 4 where the green and purple curves are not superposed. Moreover, similarly to Section 6.3.2, we

observe that the convergence in the hierarchy has not occurred at order 4 since the blue and red curves are not superimposed. As a consequence, $\hat{\mathbf{x}}_4$ is not a global minimizer of $\mathcal{J}$ but only an approximation of it. A solution to improve the quality of the minimizer is to use the solution $\hat{\mathbf{x}}_4$ as a warm start of the FB algorithm as shown by the purple curve.

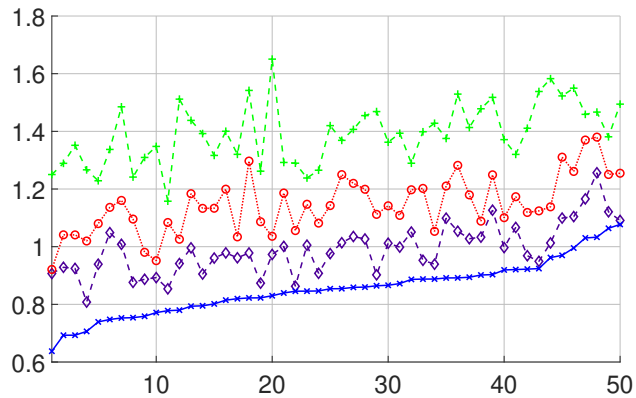


Figure 4: Comparison between the different values of the criterion for the minimizers returned by the different methods. In red $\mathcal{J}(\hat{\mathbf{x}}_4)$, in blue $\mathcal{J}_4^*$, in green $\mathcal{J}(\mathbf{x}_{\mathrm{FB0}})$, and in purple $\mathcal{J}(\mathbf{x}_{\mathrm{FB1}})$.

6.4.2. Quality of signal reconstruction

We now look at the quality of the signal reconstruction in terms of mean square error: our method is compared with several other ones to illustrate its interest for faithful recovery of the original signal $\overline{\mathbf{x}}$. In addition to the FB algorithm presented in Section 6.4.1, we compare our method with the oracle to iLASSO, a LASSO approach modified to handle the nonlinearity of the model. It consists first on applying the LASSO using a linearization of the nonlinear operator ϕ. Namely, it solves

$$\underset{\mathbf{x}\in\mathbb{R}^T}{\operatorname{argmin}} \left\|\mathbf{y} - D_\alpha(\mathcal{L}_\phi(\mathbf{h} * \mathbf{x}))\right\|^2 + \lambda_{\mathrm{LASSO}} \left\|\mathbf{x}\right\|_1 ,$$

where $\mathcal{L}_\phi$ is a linearization of ϕ and λ_{LASSO} is a parameter set empirically to 0.1. Note that for our choice of ϕ, $\mathcal{L}_\phi = \chi^{-1}$. We subsequently apply a modified iterative hard thresholding (IHT) that handles the nonlinearity. Namely, we apply the FB algorithm to find

$$\underset{\mathbf{x}\in\mathbb{R}^T}{\operatorname{argmin}} \left\|\mathbf{y} - D_\alpha(\Phi(\mathbf{h} * \mathbf{x}))\right\|^2 + \lambda_{\mathrm{IHT}} \ell_0(\mathbf{x}) ,$$

where we perform a gradient step on the data fidelity component and a proximal step on the penalization $\lambda_{\mathrm{IHT}}\ell_0$. This method provides better results than the

FB algorithm presented in Section 6.4.1. We also compare our method to the Iteratively Reweighted ℓ_1 algorithm (IRL1) [19] applied to

$$\underset{\mathbf{x}\in\mathbb{R}^T}{\operatorname{argmin}} \left\|\mathbf{y} - D_\alpha(\mathcal{L}_\phi(\mathbf{h} * \mathbf{x}))\right\|^2 + \mathcal{R}_{\lambda_{\mathrm{IRL1}}}(\mathbf{x}) \,,$$

where $\mathcal{R}_\lambda$ is the SCAD regularization. Both IRL1 and FB algorithms are initialized with the null vector.

Figure 5 illustrates the different signals for a single realization using D_2 decimation. From top to bottom, we display the original signal $\overline{\mathbf{x}}$, the subsampled observed signal $\mathbf{y}$, the signal reconstructed respectively with iLASSO $\mathbf{x}_{\mathrm{iLASSO}}$, and the signal reconstructed using our method $\hat{\mathbf{x}}_3$ at the relaxation order $k = 3$. We do not display the signal reconstructed with FB and IRL1 since those algorithms are not well suited for solving (10) and thus provide poor quality reconstruction. We first notice that iLASSO misses many peaks and also detects a peak that does not exist in the original signal while our method detects almost all peaks. One could argue the threshold coefficient λ_{IHT} in iLASSO is too high but, when we decrease it, small artifacts appear. In contrast, our method detects almost all peaks and do not leave any artifact. We observe that some peaks do not have the same amplitude as the ones in the original signal. This is due to subsampling. Indeed, if a peak is located on an even index, it will be eliminated by the subsampling. However, the convolution with h, that represents the physical limitation of sensors in our example, allows us still to recover the peak since it gets enlarged to odd neighboring. Even though, we lose information about the amplitude of this peak.

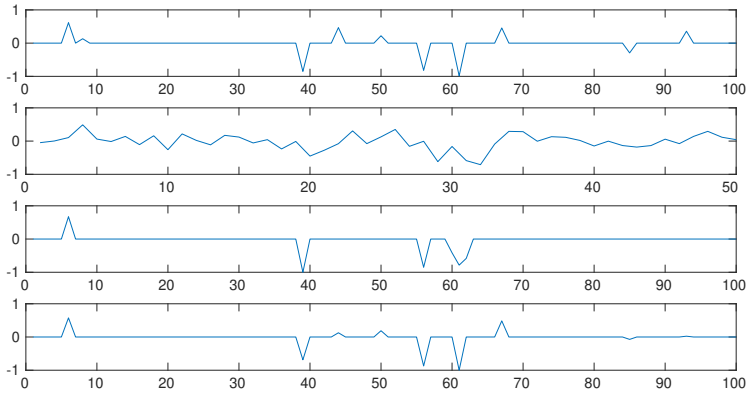
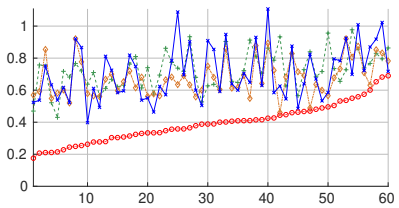
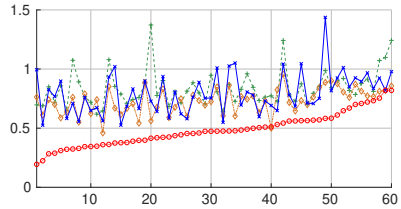


Figure 5: Comparison between iLASSO and our method for signal reconstruction under nonlinear transformation and subsampling. From top to bottom: the original signal $\overline{\mathbf{x}}$, the observed signal $\mathbf{y}$, and respectively the signal reconstructed with iLASSO $\mathbf{x}_{\mathrm{iLASSO}}$ and with our method $\hat{\mathbf{x}}_3$.

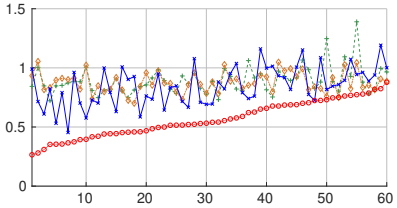
Figures 6 shows the mean square error $\|\overline{\mathbf{x}} - \mathbf{x}\| / \|\overline{\mathbf{x}}\|$ for D_∞, D_4 and D_2



(a) D_∞ decimation



(b) D_4 decimation



(c) D_2 decimation

Average values	D_∞	D_4	D_2
$\mathbf{x}_{\text{FB}}$	0.72	0.84	0.89
$\mathbf{x}_{\text{IRL1}}$	0.67	0.73	0.87
$\mathbf{x}_{\text{iLASSO}}$	0.71	0.80	0.84
$\hat{\mathbf{x}}_4$	0.39	0.48	0.56

Figure 6: Mean square error between the estimated signal and the original signal $\overline{\mathbf{x}}$. In dashed green: $\mathbf{x}_{\text{FB}}$, in dashed orange: $\mathbf{x}_{\text{IRL1}}$, in blue: $\mathbf{x}_{\text{iLASSO}}$, and in dotted red: $\hat{\mathbf{x}}_4$. Average values are shown in the table.

decimation between the original signal $\overline{\mathbf{x}}$ and: in green $\mathbf{x}_{\text{FB}}$, in orange $\mathbf{x}_{\text{IRL1}}$, in blue $\mathbf{x}_{\text{iLASSO}}$, and in red $\hat{\mathbf{x}}_4$. Those confirm the good reconstruction result shown in the specific example of Figure 5.

Finally, Table 2 shows the average computational times for different decimation operators and relaxation orders. As we expected, the better performance of our method comes at the expense of a higher computational cost than iLASSO, which takes less than 1 second.

Table 2: Computation time of our method (in seconds)

	Without oracle			With oracle		
	D_∞	D_4	D_2	D_∞	D_4	D_2
$k=2$	41	35	29	38	31	25
$k=3$	162	121	87	144	106	74
$k=4$	29991	14575	5801	24362	11062	4084

6.4.3. Handling higher-dimensional signal

Although our method provides good reconstruction results for medium-size signals, handling higher-dimensional signals is highly demanding in terms of computations as shown in our study of Section 5.1 and in Table 2. Moreover,

we observed that the memory requirements of the SDP solver for its internal process become too important. To tackle these issues, we split the signal into smaller overlapping chunks that are processed independently and then reassembled together. We illustrate the example of Figure 7 where we reconstruct a signal of dimension $T = 1000$ using 11 chunks of length 100 with 10 overlapping samples on both extremities. The overlapping sections are averaged in order to obtain the final signal. The decimation operator is D_∞ and the relaxation order is set to 3. We observe that our method yields a better reconstruction than iLASSO with a mean square error of 0.43 against 0.69 for iLASSO.

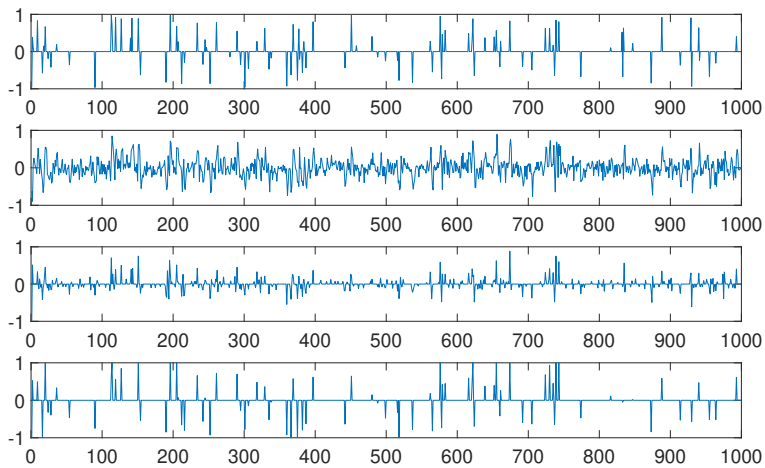


Figure 7: Reconstruction of higher-dimensional signals ($T = 1000$). From top to bottom: the original signal $\overline{\mathbf{x}}$, the observed signal $\mathbf{y}$, and respectively the signal reconstructed with iLASSO $\mathbf{x}_{\text{iLASSO}}$ and with our method $\hat{\mathbf{x}}_3$.

7. Conclusion

We have proposed a method to globally solve nonconvex problems involving exact relaxation of ℓ_0 in order to reconstruct sparse signal from degraded observations. One of the main advantages of our method is that it is able to deal with nonlinear degradations. We have first reformulated our piecewise rational criterion into a rational optimization problem before solving this problem using a hierarchy of convex SDP relaxations that benefits from the sparsity of the rational functions. We have then discussed the complexity of the obtained SDP and methods to decrease both the converging relaxation order in the hierarchy and the dimension of the SDP problem. Finally, our simulations illustrate the domain of applicability of the method and its high potential for finding a good approximation to a global minimum. Although providing good

results for medium-size problems, our method shows computational limitations for larger-scale signals and filters with longer impulse response.

References

[1] A. Marmin, M. Castella, J.-C. Pesquet, L. Duval, Signal reconstruction from sub-sampled and nonlinearly distorted observations, in: 2018 26th European Signal Processing Conference (EUSIPCO), IEEE, 2018, pp. 1970–1974. doi:10.23919/eusipco.2018.8553174.

[2] A. Marmin, M. Castella, J.-C. Pesquet, How to globally solve non-convex optimization problems involving an approximate ℓ_0 penalization, in: Proc. Int. Conf. Acoust. Speech Signal Process., IEEE, 2019, pp. 5601–5605. doi:10.1109/icassp.2019.8683692.

[3] J. Gauthier, L. Duval, J.-C. Pesquet, Optimization of synthesis oversampled complex filter banks, IEEE Trans. Signal Process. 57 (10) (2009) 3827–3843. doi:10.1109/TSP.2009.2023947.

[4] M. Q. Pham, L. Duval, C. Chaux, J.-C. Pesquet, A primal-dual proximal algorithm for sparse template-based adaptive filtering: Application to seismic multiple removal, IEEE Trans. Signal Process. 62 (16) (2014) 4256–4269. doi:10.1109/TSP.2014.2331614.

[5] C. Chaux, P. L. Combettes, J.-C. Pesquet, V. R. Wajs, A variational formulation for frame-based inverse problems, Inverse Problems 23 (4) (2007) 1495–1518. doi:10.1088/0266-5611/23/4/008.

[6] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1) (1996) 267–288.

[7] T. Blumensath, M. E. Davies, Iterative thresholding for sparse approximations, J. Fourier Anal. Appl. 14 (5-6) (2008) 629–654. doi:10.1007/s00041-008-9035-z.

[8] E. Soubies, L. Blanc-Féraud, G. Aubert, A continuous exact ℓ_0 penalty (CEL0) for least squares regularized problem, SIAM J. Imaging Sci. 8 (3) (2015) 1607–1639. doi:10.1137/151003714.

[9] M. Schetzen, Nonlinear system modelling and analysis from the Volterra and Wiener perspective, in: Lecture Notes in Control and Information Sciences, Springer London, 2010, pp. 13–24. doi:10.1007/978-1-84996-513-2_2.

[10] N. Dobigeon, J.-Y. Tourneret, C. Richard, J. C. M. Bermudez, S. McLaughlin, A. O. Hero, Nonlinear unmixing of hyperspectral images: Models and algorithms, IEEE Signal Process. Mag. 31 (1) (2014) 82–94. doi:10.1109/msp.2013.2279274.

[11] Y. Deville, L. T. Duarte, An overview of blind source separation methods for linear-quadratic and post-nonlinear mixtures, in: Latent Variable Analysis and Signal Separation, Springer International Publishing, 2015, pp. 155–167. doi:10.1007/978-3-319-22482-4_18.

[12] M. Nikolova, Description of the minimizers of least squares regularized with ℓ_0 norm. Uniqueness of the global minimizer, SIAM J. Imaging Sci. 6 (2) (2013) 904–937. doi:10.1137/11085476x.

[13] S. Bourguignon, J. Ninin, H. Carfantan, M. Mongeau, Exact sparse approximation problems via mixed-integer programming: Formulations and computational performance, IEEE Trans. Signal Process. 64 (6) (2016) 1405–1419. doi:10.1109/tsp.2015.2496367.

[14] P. L. Combettes, J.-C. Pesquet, Proximal thresholding algorithm for minimization over orthonormal bases, SIAM J. Optim. 18 (4) (2008) 1351–1376. doi:10.1137/060669498.

[15] P. L. Combettes, J.-C. Pesquet, Proximal Splitting Methods in Signal Processing, in: H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, H. Wolkowicz (Eds.), Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer, 2011, pp. 185–212. doi:10.1007/978-1-4419-9569-8.

[16] M. Castella, J.-C. Pesquet, Optimization of a Geman-McClure like criterion for sparse signal deconvolution, in: Proc. IEEE Int. Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), IEEE, Cancun, Mexico, 2015, pp. 309–312. doi:10.1109/camsap.2015.7383798.

[17] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Am. Stat. Assoc. 96 (456) (2001) 1348–1360. doi:10.1198/016214501753382273.

[18] P. Ochs, A. Dosovitskiy, T. Brox, T. Pock, On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision, SIAM J. Imaging Sci. 8 (1) (2015) 331–372. doi:10.1137/140971518.

[19] E. J. Candès, M. B. Wakin, S. P. Boyd, Enhancing sparsity by reweighted ℓ_1 minimization, J. Fourier Anal. Appl. 14 (5-6) (2008) 877–905. doi:10.1007/s00041-008-9045-x.

[20] P. Breheny, J. Huang, Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection, Ann. Appl. Stat. 5 (1) (2011) 232–253. doi:10.1214/10-aoas388.

[21] A. Patrascu, I. Necoara, Random coordinate descent methods for ℓ_0 regularized convex optimization, IEEE Trans. Automat. Contr. 60 (7) (2015) 1811–1824. doi:10.1109/tac.2015.2390551.

[22] I. Selesnick, Sparse regularization via convex analysis, IEEE Trans. Signal Process. 65 (17) (2017) 4481–4494. doi:10.1109/tsp.2017.2711501.

[23] T. Zhang, Analysis of multi-stage convex relaxation for sparse regularization, J. Mach. Learn. Res. 11 (2010) 1081–1107.

[24] C.-H. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Appl. Stat. 38 (2) (2010) 894–942. doi:10.1214/09-aos729.

[25] M. Artina, M. Fornasier, F. Solombrino, Linearly constrained nonsmooth and nonconvex minimization, SIAM J. Optim. 23 (3) (2013) 1904–1937. doi:10.1137/120869079.

[26] A. Jezierska, H. Talbot, O. Veksler, D. Wesierski, A fast solver for truncated-convex priors: Quantized-convex split moves, in: Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2011, pp. 45–58. doi:10.1007/978-3-642-23094-3_4.

[27] J. B. Lasserre, Global optimization with polynomials and the problem of moments, SIAM J. Optim. 11 (3) (2001) 796–817. doi:10.1137/s1052623400366802.

[28] M. Castella, J.-C. Pesquet, A. Marmin, Rational optimization for nonlinear reconstruction with approximate ℓ_0 penalization, IEEE Trans. Signal Process. 67 (6) (2019) 1407–1417. doi:10.1109/tsp.2018.2890065.

[29] C. Vendeuvre, R. Ruiz-Guerrero, F. Bertoncini, L. Duval, D. Thiébaut, M.-C. Hennion, Characterisation of middle-distillates by comprehensive two-dimensional gas chromatography (GC $\times$ GC): A powerful alternative for performing various standard analysis of middle-distillates, J. Chromatogr. A 1086 (1-2) (2005) 21–28. doi:10.1016/j.chroma.2005.05.106.

[30] C. Vendeuvre, R. Ruiz-Guerrero, F. Bertoncini, L. Duval, D. Thiébaut, Comprehensive two-dimensional gas chromatography for detailed characterisation of petroleum products, Oil Gas Sci. Tech. 62 (1) (2007) 43–55. doi:10.2516/ogst:2007004.

[31] A. Felinger (Ed.), Data analysis and signal processing in chromatography, Elsevier, 1998.

[32] Y. Kalambet, Y. Kozmin, A. Samokhin, Comparison of integration rules in the case of very narrow chromatographic peaks, Chemometr. Intell. Lab. Syst. 179 (2018) 22–30. doi:10.1016/j.chemolab.2018.06.001.

[33] E. Soubies, L. Blanc-Féraud, G. Aubert, A unified view of exact continuous penalties for $\ell_2-\ell_0$ minimization, SIAM J. Optim. 27 (3) (2017) 2034–2060. doi:10.1137/16m1059333.

[34] J. B. Lasserre, Moments, Positive Polynomials and Their Applications, Imperial College Press, London, U.K., 2009.

[35] D. Henrion, Optimization on linear matrix inequalities for polynomial systems control (Sep. 2013). arXiv:1309.3112v1.

[36] J. Nie, Optimality conditions and finite convergence of Lasserre's hierarchy, Math. Programm. 146 (1-2) (2013) 97–121. doi:10.1007/s10107-013-0680-x.

[37] D. Henrion, J.-B. Lasserre, Detecting global optimality and extracting solutions in GloptiPoly, in: Positive Polynomials in Control, Vol. 312, Springer Berlin Heidelberg, 2005, pp. 293–310. doi:10.1007/10997703_15.

[38] H. Waki, S. Kim, M. Kojima, M. Muramatsu, Sums of squares and semidefinite program relaxations for polynomial optimization problems with structured sparsity, SIAM J. Optim. 17 (1) (2006) 218–242. doi:10.1137/050623802.

[39] F. Bugarin, D. Henrion, J. B. Lasserre, Minimizing the sum of many rational functions, Math. Program. Comput. 8 (1) (2015) 83–111. doi:10.1007/s12532-015-0089-z.

[40] D. Henrion, J.-B. Lasserre, J. Löfberg, GloptiPoly 3: moments, optimization and semidefinite programming, Optim. Methods Softw. 24 (4-5) (2009) 761–779. doi:10.1080/10556780802699201.

[41] K. C. Toh, M. J. Todd, R. H. Tütüncü, SDPT3 — a Matlab software package for semidefinite programming, version 1.3, Optim. Methods Softw. 11 (1-4) (1999) 545–581. doi:10.1080/10556789908805762.

A. Detailed computation complexity of relaxed SDP problems

We detail here the computation of the complexity of an SDP problem, i.e. the quadruple $(n, m, m_{\mathrm{s}}, \ell)$, depending on the initial data like U, L and T as well as the relaxation order k.

A.1. Number of blocks m_{s}

In order to solve (13), we introduce $U+T$ measures. Moment and localizing matrices of each measure yield a block in the SDP problems of the hierarchy. There is one moment matrix per measure, i.e. a total of $U+T$ moment matrices. The number of localizing matrices for each measure is equal to the number of polynomial constraints defining the set $\mathcal{K}_i$. Equation (26) gives T_i constraints for the definition of each set $\mathcal{K}_i$ associated to the measures $(\mu_u)_{u\in[\![1,U]\!]}$ while (27) gives $1+3I$ constraints for each set $\mathcal{K}_i$ associated to the measures $(\nu_t)_{t\in[\![1,T]\!]}$. Indeed the polynomial equality constraint in (27) is translated into two polynomial inequality constraints. The first $L-1$ measures $(\mu_u)_{u\in[\![1,U]\!]}$ are defined on a number T_i of variables smaller than L due to the convolution filter. In the following, we neglect it for the sake of clarity and assume that T_i is equal to L for all the measure $(\mu_u)_{u\in[\![1,U]\!]}$. Thus, the final number of blocks in the matrices $\mathbf{C}$ and $(\mathbf{A}_i)_{i\in[\![1,m]\!]}$ in (21) is

$$m_{\mathrm{s}} = U(1+L) + T(2+3I)\,.$$

It is interesting to notice that the relaxation order k does not have any effect on the number of blocks; it only increases the size of the blocks.

A.2. Number of linear equality constraints ℓ

We then count the number of linear equality constraints in (25), without considering the redundant ones. For u belonging to $[\![1, U-1]\!]$, θ_u denotes the overlap parameter defined as the number of variables shared between g_u and g_{u+1}. Note that θ_u depends on u but also on the length of the filter L and on the parameter of the decimation δ. Furthermore, we remark that all the rational functions $(g_u)_{u\in[\![1,U]\!]}$ have same degree at their numerator and denominator. We denote their denominator by q_u, and we define $\mathrm{d}_q = \mathrm{d}_{q_u}$.

Following Section 4.3.2, we need to consider equality of moments of monomials in θ_u variables up to degree $2(k-d_q)$, which gives $\binom{\theta_u+2(k-d_q)}{2(k-d_q)}$ equality constraints for every u in $[\![1, U-1]\!]$ on consecutive measures $(\mu_u)_{u\in[\![1,U]\!]}$. Adding the linear constraints linking moment related to $(\mu_u)_{u\in[\![1,U]\!]}$ and $(\nu_t)_{t\in[\![1,T]\!]}$, we finally obtain

$$\ell = \sum_{u=1}^{U-1} \binom{\theta_u + 2(k-d_q)}{2(k-d_q)} + 2(k-d_\zeta)T\,,$$

where d_ζ corresponds to (1) for the maximal degree of the denominator of rational function $(\zeta_i)_{i\in[\![1,I]\!]}$. The impact of linear equality constraints on the computational time of SDP solver is minor compared to n, m and m_s.

A.3. Dimension of the global moment vector m

The dimension m of the vector $\mathbf{v}$ is simply obtained by summing up the dimension of the moment vectors for all the measures $(\mu_u)_{u\in[\![1,U]\!]}$ and $(\nu_t)_{t\in[\![1,T]\!]}$. Considering $(\mu_u)_{u\in[\![1,U]\!]}$ as U measures on L variables and $(\nu_t)_{t\in[\![1,T]\!]}$ as T measures on $1+I$ variables, it follows that

$$m = U\binom{L+2k}{2k} + T\binom{1+I+2k}{2k}.$$

A.4. Dimension of the semi-definite constraint n

At last, n is the sum of all the block sizes of the matrices in the SDP problem, that is the sum of the size of the all moment and localizing matrices. The U moment matrices corresponding to measures $(\mu_u)_{u\in[\![1,U]\!]}$ have size $\binom{L+k}{k}$ while the ones corresponding to $(\nu_t)_{t\in[\![1,T]\!]}$ have size $\binom{1+I+k}{k}$. Since all the polynomial constraints defining the sets $(\mathcal{K}_i)_{i\in[\![1,U+T]\!]}$ are linear or quadratic, the localizing matrices have respectively a size of $\binom{L+k-1}{k-1}$ for measures $(\mu_u)_{u\in[\![1,U]\!]}$ and $\binom{I+k}{k-1}$ for measures $(\nu_t)_{t\in[\![1,T]\!]}$. Finally, we obtain

$$n = U\left(\binom{L+k}{k} + L\binom{L+k-1}{k-1}\right) + T\left(\binom{1+I+k}{k} + (1+3I)\binom{I+k}{k-1}\right).$$