HAL open science

Entropic Lower Bound of Cardinality for Sparse Optimization

Quentin Jacquet, Agnes Bialecki, Laurent El Ghaoui, Stéphane Gaubert, Riadh Zorgati

To cite this version:

Quentin Jacquet, Agnes Bialecki, Laurent El Ghaoui, Stéphane Gaubert, Riadh Zorgati. Entropic Lower Bound of Cardinality for Sparse Optimization. 2024. hal-03874638v2

HAL Id: hal-03874638

https://hal.science/hal-03874638v2

Preprint submitted on 3 Apr 2024

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.

Entropic Lower Bounds of ℓ_0-*pseudo-norm* for Sparse Optimization

Quentin Jacquet[1,2], Agnès Bialecki[2], Laurent El Ghaoui[3], Stéphane Gaubert[1], Riadh Zorgati[2]

[1] INRIA, CMAP, Ecole Polytechnique, Palaiseau, France

stephane.gaubert@inria.fr

[2] EDF R&D Saclay, Palaiseau, France

{quentin.jacquet,riadh.zorgati}@edf.fr

[3] VinUniversity, Vietnam

laurent.eg@vinuni.edu.vn

Abstract

We introduce a family of cardinality's lower bounds, defined as ratios of norms. We prove that the tightest bound of the family is obtained as a limit case, and involves a Shannon entropy. We then use this entropic lower bound in sparse optimization problems to approximate cardinality requirements. This provides a nonlinear nonconvex relaxed problem, which can be efficiently solved by off-the-shelf nonlinear solvers. In the numerical study, we focus on the case where the optimization is performed on the simplex, and where the classical ℓ_1 penalization does not yield sparse solution. The Finance Index Tracking problem is taken as an example and illustrates the efficiency of the proposed approach.

Keywords: Sparse Optimisation, Cardinality, ℓ_0-pseudo-norm, Shannon entropy, Index Tracking.

1 Introduction

In numerous fields such as finance, energy or machine learning, decision makers aim to control the cardinality of the solution vector, i.e., the number of representative features (assets in portfolio optimization Bertsimas and Shioda, 2007, shutdowns/start-ups in thermal power plants scheduling Bialecki et al., 2014a, Support Vectors in machine learning Bi et al., 2003, ...).

In optimization terminology, the cardinality of a solution encoded by a vector $x \in \mathbb{R}^n$ is the number of non-zero elements, i.e., $|\{i \in [n] : x_i \neq 0\}|$, and is often written $\mathrm{card}(x)$. Both correspond to the so-called ℓ_0-*pseudo-norm*, denoted $\|x\|_0$. This pseudo-norm is positively homogeneous of degree 0, meaning that for all $x \in \mathbb{R}^n$ and $\alpha \neq 0$, $\|\alpha x\|_0 = \|x\|_0$. Optimization under cardinality requirements is called *sparse optimization*.

Sparsity has mainly emerged from the signal processing and machine learning communities, under names such as compressed sensing Donoho, 2006 and sparse learning Bi et al., 2003. In machine learning, the Sparse Support Vector Machine aims at finding a minimal cardinality linear classifier which can separate two classes of labeled data. The sparsity of the solution helps for better interpretability of the solution which is crucial in automated analysis of large text corpora. One major case of using sparsity is the feature selection which refers to the necessity of selecting representative variables

from datasets containing a large number of features, many of them being irrelevant or redundant. For example, in finance, feature selection is used to restrict asset allocation to a limited number of assets in the portfolio. Sparsity allows reducing a priori the dimension of a large-scale problem when performing a sparse regression that may be more efficient than the classical one, by selecting a small set of predictors in a least-squares sense.

Sparsity is also very useful in energy management where many problems involve cardinality constraints. Our original motivation is the intra-day problem, consisting in updating a day-ahead generation schedule by modifying a limited number of power units schedules (Bialecki et al., 2014a, Bialecki et al., 2014b). Two other examples concern operation of power plants. During start-up, some components of thermal power plants go from 20°C to $1300 - 1900$°C in a few seconds leading, over the long term, to damages, reducing their lifespan. Saving durability of these plants consists in limiting the number of shutdowns/start-ups. Finally, when operating nuclear power plants, it is necessary to limit the number of "deep" drops in power (because a nuclear reaction at low power for a long time generates unwanted isotopes that "poison the heart") and also to limit the daily number of production variations (modulations) so as not to over-consume the boron (neutron absorber) because a reduction in the boron available in the core makes the plant more difficult to operate and leads to its premature shutdown for refueling.

Optimization problems involving the ℓ_0-norm of the decision vector belong to the class of sparse optimization problems and take one of the two following general form:

(i) $\|x\|_0$ in the objective function:

$$\min_{x \in X} \{f(x) + \lambda \|x\|_0 \ \mid \ g(x) \leq 0\} \tag{P_λ}$$

(ii) $\|x\|_0$ in constraints:

$$\min_{x \in X} \{f(x) \ \mid \ g(x) \leq 0, \ \|x\|_0 \leq k \ (< n)\} \tag{P_k}$$

In both formulations, $X \subseteq \mathbb{R}^n$ is the set defining the constraints. The objective function f corresponds to a given criterion and is often considered as convex in machine learning applications (such as *least-squares problems* (LSQ)) while it may be nonconvex in energy applications. The parameter $\lambda \geq 0$ is viewed as a regularization parameter used to manage the trade-off between the criterion $f(x)$ and the sparsity of x. In selection problems, a stronger constraint on the decision vector arises: x must belongs to the probability simplex, i.e., $X \subseteq \Delta_n := \left\{x \in \mathbb{R}^n_+ \mid \sum_{i=1}^n x_i = 1\right\}$. In this specific case, the ℓ_1-norm is constrained to be one.

Since the $\|\cdot\|_0$ is lower semicontinuous on $\mathbb{R}^n$ and is discontinuous at any point belonging to an hyperplane $x_i = 0$, optimization problems involving the ℓ_0-norm are nonconvex and hence very challenging. They are inherently of combinatorial nature and hence, not solvable in polynomial-time in general **Bienstock'1996**. Huge research effort has been made in sparse optimization and several approaches have been proposed. Let us cite :

- **The convex approximation**. A typical example is the famous Least Absolute Shrinkage and Selection Operator (LASSO) penalty technique. It consists in replacing the nonconvex term $\|x\|_0$ by the convex approximation $\|x\|_1$. This approach has first been proposed for linear regression in Tibshirani, 1996. Since then, the ℓ_1-regularization technique has been extensively studied and improved (Gribonval and Nielsen, 2003, Zou, 2006, Knight and W. Fu, 2000, ...) This leads to

very efficient and scalable algorithms in many cases. For example, the main approaches to sparse learning replace the (hard) cardinality requirements with some simpler (convex) functions such as the ℓ_1-norm, leading to tractable optimization problems. However, in several applications of great interest, in energy for instance, the solutions obtained in this way are generally far from the expected one. Moreover, replacing cardinality by the convex approximation based on ℓ_1-norm is pointless for optimization problems over the probability simplex (selection problems) i.e., when the variables are discrete probability distributions, since in this case the ℓ_1 norm is constant over the feasible set. Then, the now-standard approaches fail and some methods have been specifically dedicated to sparse optimization on simplex, finding alternative convex approximations, for e.g. based on the ℓ_∞-norm Pilanci et al., 2012.

- **The nonconvex approximation**. This approach consists in approximating $\|x\|_0$ by a continuous nonconvex function. Various functions have been proposed to approximate the ℓ_0 term (Bradley and Mangasarian, 1998, W.J. Fu, 1998, Weston et al., 2003) and several types of algorithms have been designed to solve related optimization problems, including algorithms based on the Difference of Convex functions (DC) (Chen et al., 2010, Gasso et al., 2009, Guan and Gray, 2013, Ong and Le Thi, 2013, Thiao et al., 2008, Pham Dinh and Le Thi, 2014) or based on Successive or Local Linear Approximation (Bradley and Mangasarian, 1998, Zou and Li, 2008). Nonconvex approximations can be better than convex relaxations by guarantying a higher sparsity level, but the related nonconvex optimization problems are more difficult to solve.

- **Heuristic approach**. In addition to the mathematical programming based approaches, heuristic methods have also been applied, especially greedy algorithms, designed to directly tackle cardinality minimization problem. Two noteworthy examples are the matching pursuit Mallat and Zhang, 1993 and the orthogonal matching pursuit Pati et al., 1993.

Table 1 gives some additional entries in the literature.

	Problem	Optimality	Resolution
Bertsimas and Shioda, 2007	Sparse LSQ Portfolio selection	Global	Branch & Bound
Nadisic et al., 2020	Sparse LSQ	Global	Branch & Bound
Ben Mhenni et al., 2021	Sparse LSQ	Global	Branch & Bound
Tibshirani, 1996	Sparse LSQ	Relaxation	Convex penalization (LASSO)
Soubies et al., 2015	Sparse LSQ	Relaxation	Continuous nonsmooth penalty
Haddou and T., 2019	Sparsity	Relaxation	Nonconvex penalization
Atamturk and Gomez, 2019	Sparse regression	Lower Bound	SDP (convex)
Chancelier and De Lara, 2019	GSO	Lower Bound	Caprac conjugacy
Soussen et al., 2011	Sparse LSQ	Heuristic	Penalization + Greedy

Table 1: Different approaches to sparse optimization.
LSQ: Least squares problem
GSO : General Sparse Optimization

In this context, we propose an approach based on constructing a set of lower bounds of ℓ_0-pseudo-norm expressed as ratios of norms (Theorem 2.1). In particular, we prove that the best lower bound we obtained is expressed as a function of Shannon entropy Shannon, 1948 and ℓ_1-norm. In Sakai and Iwata, 2016, the authors bring to light sharp extreme relations between Shannon entropy and ℓ_α-norm ($\alpha > 0$). Here, we obtain a relation for $\alpha = 0$. Then, we insert this new bound in sparse optimization problems, and show that the relaxed problem is a smooth nonlinear problem (yet non

convex), see Proposition 2.2. Then, a local solution can be obtained by using a nonlinear solvers like IPOPT Wächter and Biegler, 2006. Numerical experiments on the Finance Index Tracking problem illustrate the efficiency of the proposed approach (Section 4).

2 Entropic Lower Bound of $\|x\|_0$ and use in Sparse Optimization

2.1 Renyi's entropies

Recall that the Renyi's entropy Rényi et al., 1961 of order $\alpha \geq 0, \alpha \neq 1$, associated to a discrete distribution $p \in \mathbb{R}^n$, $p \geq 0$, $p_1 + \cdots + p_n = 1$, is the quantity:

$$H_\alpha(p) := \left(\frac{1}{1-\alpha}\right) \log \sum_{i=1}^{n} p_i^\alpha .$$

Depending on the value of parameter α, four important special cases of Renyi's entropies can be mentioned:

- ⋄ Hartley's entropy Hartley, 1928 ($\alpha = 0$): $H_0(p) = \log \|x\|_0$.
- ⋄ Shannon's entropy Shannon, 1948 ($\alpha \to 1$): $H_1(p) = \lim_{\alpha \to 1} H_\alpha(p) = -\sum_{i \in [n]} p_i \log p_i$.
- ⋄ Collision entropy ($\alpha = 2$): $H_2(p) = -\log \sum_{i \in [n]} p_i^2 = -\log \|p\|_2^2$.
- ⋄ Minimal entropy ($\alpha \to \infty$): $H_\infty(p) = \lim_{\alpha \to \infty} H_\alpha(p) = -\log \|p\|_\infty$.

In the case of a uniform probability distribution, the Rényi entropies of all orders, the Hartley's entropy and the Shannon entropy coincide.

The natural logarithm of ℓ_0-*pseudo-norm* of a vector $x \in \mathbb{R}^n$ is the Hartley's entropy, a measure of uncertainty Hartley, 1928, corresponding to the information provided by selecting, randomly and uniformly, a sample from x.

2.2 A hierarchy of lower bounds

We define the ℓ_q-normof a vector $x \in \mathbb{R}^n$, $p \geq 1$, as:

$$\|x\|_q = \left(\sum_{i=1}^{n} |x_i|^q\right)^{\frac{1}{q}} .$$

We remind the known lower bounds of $\|x\|_0$ as ratios of norms ($\forall x \in \mathbb{R}^n \setminus \{\mathbf{0}\}$):

$$B_\infty(x) := \frac{\|x\|_1}{\|x\|_\infty} \leq \|x\|_0 \tag{1}$$

$$B_2(x) := \left(\frac{\|x\|_1}{\|x\|_2}\right)^2 \leq \|x\|_0 . \tag{2}$$

These lower bounds may be far from $\|x\|_0$ in practice.

We now introduce a family of bounds generalizing the two previous bounds: for $x \neq 0$, and $\alpha > 0$, define

$$B_\alpha(x) := \left(\frac{\|x\|_1}{\|x\|_\alpha}\right)^{\frac{\alpha}{\alpha-1}} = \exp H_\alpha(p(x)) = \left(\sum_{i\in[n]} p_i(x)^\alpha\right)^{\frac{1}{\alpha-1}} \quad , \quad p(x) := |x|/\|x\|_1 .$$

In particular,

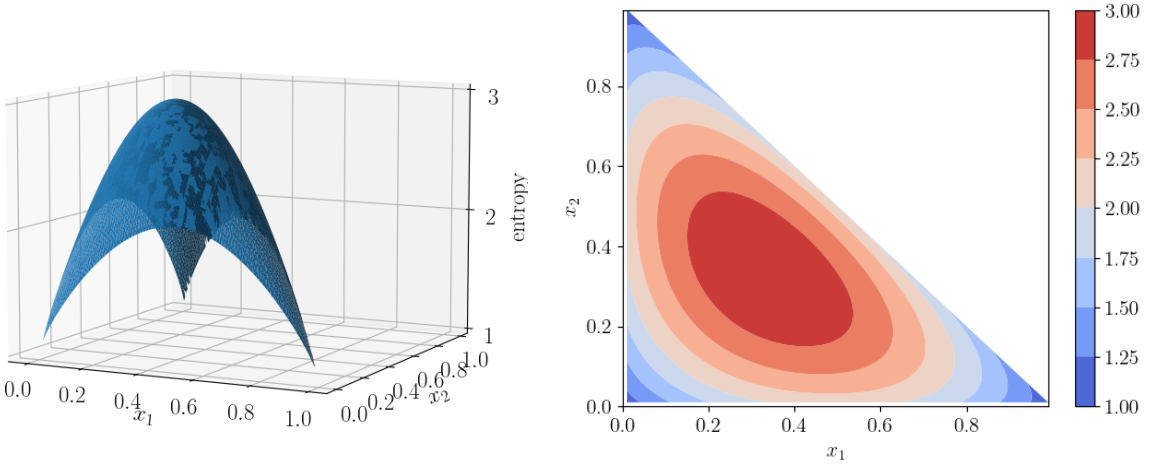
$$B_1(x) = \frac{\|x\|_1}{\prod_{i\in[n]} |x_i|^{|x_i|/\|x\|_1}} = \|x\|_1 \exp\left(-\frac{1}{\|x\|_1}\sum_{i\in[n]} |x|_i \log |x|_i\right) \quad . \tag{3}$$

Theorem 2.1 recalls that the family $(B_\alpha)_{\alpha\in]0,+\infty[}$ is ordered in a decreasing fashion, so that the quality of the bound improves when α decreases.

Theorem 2.1 (Monotonicity according to order α, see e.g. Cachin, 1997)**.**

$$B_\infty(x) \leq \cdots \leq B_2 \leq \cdots \leq B_1 \leq \cdots \leq B_0 = \|x\|_0 \quad . \tag{4}$$

In the case $\|x\|_1 = 1$, B_1 simplifies to the exponential of the Shannon entropy. We refer to Figure 1 for a numerical example of the bound B_1. This illustrates, in particular, the concavity of this nonlinear bound.



(a) Entropy in 3 dimensions

(b) Entropy as a contour plot

Figure 1: Shannon entropy $H_1(x)$ for $x \in \Delta_3$. The two first dimensions x_1 and x_2 are displayed, and the third one is implicitly defined as $x_3 = 1 - x_1 - x_2$.

2.3 Sparse optimization and focus on Shannon entropy

We now focus on the integration of the previously defined entropic bound (3) in a sparse optimization problem: let us assume a generic problem of the form (P_k). The corresponding relaxation is then

$$\begin{aligned} \min_{x\in X} \quad & f(x) \\ \text{s.t.} \quad & g(x) \leq 0 \\ & B_1(x) \leq k \end{aligned} \tag{$\widetilde{P}_k$}$$

Proposition 2.2. *The problem* $(\widetilde{P}_k)$ *can be equivalently reformulated as*

$$\begin{aligned} \min_{x\in X} \quad & f(x) \\ \text{s. t.} \quad & g(x) \leq 0 \\ & \Gamma(x, \|x\|_1) \leq 0 \end{aligned} \tag{5}$$

where

(i) $\Gamma : (x_1, \ldots, x_n, z) \in \mathbb{R}_+^{n+1} \mapsto z \log(z) - \sum_{i\in[n]} x_i \log(kx_i)$,

(ii) the Jacobian of Γ *is defined as* $\frac{\partial \Gamma}{\partial x_i}(x, z) = -1 - \log(kx_i)$, $\frac{\partial \Gamma}{\partial z}(x, z) = 1 + \log(z)$,

(iii) the Hessian of Γ *is* $H_\Gamma \coloneqq \operatorname{diag}(-1/x_1, \ldots, -1/x_n, 1/z)$ *for* $(x, z) \in \mathbb{R}_{>0}^{n+1}$.

The proof is immediate.

The relaxation problem that we obtain is not convex (the function Γ is concave), and there is no guarantee in finding the global optimum of this relaxation. Nonetheless, this problem numerically leads to solutions which are both sparse and with satisfactory objective value, see Section 4.

3 Metric estimates between B_α and ϵ-cardinality

3.1 Majorization and Schur-convexity

Definition 3.1 (Majorization)**.** *For a vector* $a \in \mathbb{R}_+^n$*, we denote by* $a^\downarrow \in \mathbb{R}_+^n$ *the vector with the same components, but sorted in descending order. Given* $a, b \in \mathbb{R}_+^n$*, we say that* a *weakly majorizes (or dominates)* b *from below written* $a \succ_w b$ *iff*

$$\sum_{i=1}^k a_i^\downarrow \geq \sum_{i=1}^k b_i^\downarrow \quad \text{for } k = 1, \ldots, n \ .$$

If $a \succ_w b$ *and in addition* $\sum_{i=1}^n a_i = \sum_{i=1}^n b_i$*, then we say that* a *majorizes* b*, written* $a \succ b$.

Definition 3.2 (Schur-convexity/concavity)**.** *Let* $\mathcal{A} \subset \mathbb{R}_+^n$*. A real-valued function* $\phi : \mathbb{R}_+^n \to \mathbb{R}$ *is said to be* Schur-convex *(resp.* Schur-concave*) if* $\phi(x) \leq \phi(y)$ *(resp.* $\phi(x) \geq \phi(y)$ *for any* $x, y \in \mathcal{A}$ *satisfying* $x \prec y$.

Proposition 3.3 (Marshall et al., 2011, Appendix F.3.a (p.532))**.** *The Rényi entropy of an arbitrary* $\alpha > 0$ *is Schur-concave; in particular, for* $\alpha = 1$*, the Shannon entropy is Schur-concave.*

3.2 Extreme relation between ϵ-cardinality and entropy

We first show that no tight relation can be found between the cardinality and the bound B_α. To see this, let $k < n$, and $\epsilon < 1/n$. Define the probability distribution $v_n(k, \epsilon) \in \Delta_n$ as

$$[v_n(k, \epsilon)]_i = \begin{cases} 1 - (k-1)\epsilon, & i = 1 \\ \epsilon, & 2 \leq i \leq k \\ 0, & k+1 \leq i \leq n \end{cases} \tag{6}$$

and the associated entropy

$$H_\alpha^v(k,\epsilon) \coloneqq H_\alpha(v_n(k,\epsilon)) = \frac{1}{1-\alpha}\ln\left((1-(k-1)\epsilon)^\alpha + (k-1)\epsilon^\alpha\right), \qquad 0<\alpha<1$$
$$H_\alpha^v(k,\epsilon) \coloneqq H_\alpha\left(v_n(k,\epsilon)\right) = -\left(1-(k-1)\epsilon\right)\ln\left(1-(k-1)\epsilon\right) - (k-1)\epsilon\ln(\epsilon), \qquad \alpha=1$$

The following property is immediate.

Proposition 3.4 (Worst-case comparison between cardinality and B_α)**.** *For any $\epsilon > 0$ and $0 < \alpha \leq 1$, $\operatorname{card}(v_n(n,\epsilon)) = n$ and $B_\alpha(v_n(n,\epsilon)) \xrightarrow[\epsilon\to 0]{} 1$. Therefore, the cardinality of a given probability distribution $p \in \Delta_n$ is not* controlled *by the estimation $B_\alpha(p)$.* □

Nonetheless, we aim to find extreme relations between B_α and the ϵ-*cardinality*, defined as

$$\operatorname{card}_\epsilon(p) = |\{i \in [n] \mid p_i \geq \epsilon\}| \quad . \tag{7}$$

The parameter ϵ is viewed as a filtering threshold.

Lemma 3.5. *Viewing k as a real number, the function $k \in [1,n] \mapsto H_\alpha^v(k,\epsilon)$ is an increasing function for $\epsilon \leq \frac{1}{n}$ and $0 < \alpha \leq 1$.*

Proof. For $\alpha = 1$, $\frac{\partial H_\alpha^v}{\partial k}(k,\epsilon) = \epsilon\left[1+\ln\left(\frac{1}{\epsilon}-k+1\right)\right]$. As $\epsilon \leq \frac{1}{n}$ and $k \leq n$, we get that $k \mapsto H_v(k,\epsilon)$ is increasing. Now, for $0 \leq \alpha < 1$, $\frac{\partial}{\partial k}\exp\left((1-\alpha)H_\alpha^v(k,\epsilon)\right) = \epsilon^\alpha - \alpha\epsilon(1-(k-1)\epsilon)^\alpha \geq \epsilon^\alpha - \epsilon > 0$. □

Lemma 3.6. *For any $\epsilon > 0$ and $0 < \alpha \leq 1$, an optimal solution of the problem*

$$\min_{p\in\Delta_n}\{H_\alpha(p) \mid \operatorname{card}_\epsilon(p) = k\} \tag{$P_{\alpha,\epsilon}^{k,n}$}$$

is $v_n(k,\epsilon)$, and corresponds to an objective value $H_\alpha^v(k,\epsilon)$.

Proof. Any ordered element $p \in \Delta_n$ satisfying $\operatorname{card}_\epsilon(p) = k$ can be represented as

$$p = \left(1 - \sum_{i=1}^{k-1}\alpha_i - \sum_{i=k}^{n}\beta_i, \ldots, \alpha_{k-1}, \beta_k, \ldots, \beta_n\right) \quad ,$$

with $\alpha_1 \geq \cdots \geq \alpha_{k-1} \geq \epsilon$ and $\epsilon > \beta_k \geq \cdots \geq \beta_n \geq 0$. Then, for $1 \leq d \leq n$,

$$\sum_{i=1}^{d}[v_n(k,\epsilon)]_i - \sum_{i=1}^{d}p_i = \begin{cases} \sum_{i=d}^{k-1}\alpha_i - (k-d)\epsilon + \sum_{i=k}^{n}\beta_i, & d \leq k \\ \sum_{i=d}^{n}\beta_i, & d > k \end{cases} .$$

By using Proposition 3.3, we obtain that the minimum of the Rényi entropy is attained for $v_n(k,\epsilon)$. □

Finding the distribution giving the minimal Rényi entropy using majorization theory has been also performed in Koga, 2013 and Sason, 2018 for different set of constraints. Also, extreme relations between Rényi entropy and l_q-norm, $q > 0$, have been found in Sakai and Iwata, 2016.

We introduce the invertible, increasing, function $\phi_{\alpha,\epsilon} : k \in [1,n] \mapsto \exp H_\alpha^v(k,\epsilon) \in [1,n]$.

Theorem 3.7 (ϵ-cardinality bounds)**.** *Let $1 \leq b \leq n$, $\epsilon > 0$ and $0 \leq \alpha \leq 1$. For any vector $p \in \Delta_n$, if $B_\alpha(p) \leq b$, then $\operatorname{card}_\epsilon(p) \leq \lfloor \phi_{\alpha,\epsilon}^{-1}(b) \rfloor$.*

Proof. By the resolution of $(P^{k,n}_{\alpha,\epsilon})$ (Lemma 3.6), we know that

$$\mathrm{card}_\epsilon(p) = k \Rightarrow B_\alpha(p) \geq \exp H^v_\alpha(k,\epsilon)$$

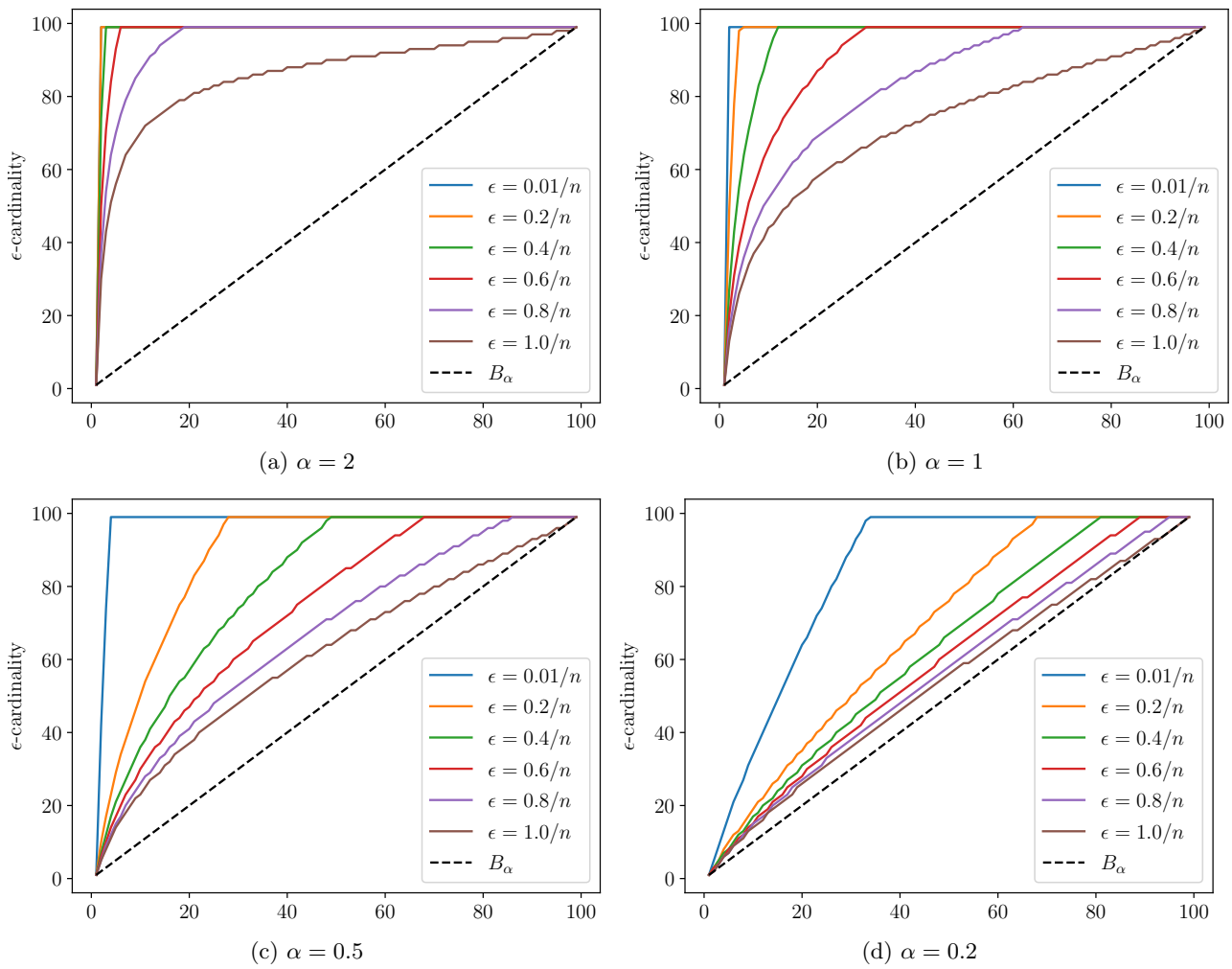
As $\phi_{\alpha,\epsilon}$ is increasing and invertible, we deduce that $\mathrm{card}_\epsilon(p) \geq k \Rightarrow B_\alpha(p) \geq \exp H^v_\alpha(k,\epsilon)$ and so

$$B_\infty(p) \leq b \Rightarrow \mathrm{card}_\epsilon(p) \leq \phi^{-1}_{\alpha,\epsilon}(b) \ .$$

□

Remark 3.1. *The relation found in Theorem 3.7 is tight as it is attained for* $p = v_n(\phi^{-1}_{\alpha,\epsilon}(b), \epsilon)$ *if* $\phi^{-1}_{\alpha,\epsilon}(b) \in \mathbb{N}$.

Theorem 3.7 provides sparsity guarantees for the solution. In fact, if one requires a maximum cardinality of b, the solution has an ϵ-cardinality of $\lfloor \phi^{-1}_{\alpha,\epsilon}(b) \rfloor$. Figure 2 shows that the tightness of the bound improves when ϵ grows and α decreases.



(a) $\alpha = 2$ (b) $\alpha = 1$

(c) $\alpha = 0.5$ (d) $\alpha = 0.2$

Figure 2: The ϵ-cardinality upper bound $b \mapsto \lfloor \phi^{-1}_{\alpha,\epsilon}(b) \rfloor$ for $1 \leq b \leq n = 100$. The approximation becomes tighter when ϵ increases and when α decreases.

4 Numerical Experiments

As an illustration of our approach, we will consider a sparse regression problem on the simplex with a use case from finance (Index tracking).

A financial index is a number representing the value of the set of assets (stocks or bonds) which reflects the value of a specific market or a segment of it. Insofar as an index is not a financial instrument that we can directly trade, a stock or a bond market index is effectively equivalent to a hypothetical portfolio of assets. In order to gain access to an index, it is necessary to use financial instruments such as options, futures and exchange-traded funds, or to create a portfolio of assets that closely tracks a given index. For a given index, fund managers have the choice between two basic investment strategies. The active strategy assumes that the markets are not perfectly efficient so that fund managers, thanks to their know-how, makes specific investments and hope to add value by choosing high performing assets outperforming an investment benchmark index. On the contrary, the passive strategy assumes that the market cannot be beaten in the long run, so that fund managers expect a return that closely replicates the investment weighting and returns of a benchmark index.

Currently, passive strategies seem to attract more interest from investors. Index tracking, also known as index replication, is one of the most popular passive portfolio management strategy to use the market index to determine the portfolio weights by reproducing the performance of a market index, i.e., to match the performance of a theoretical portfolio as closely as possible. Index tracking allows to get the desired returns from the overall market growth with the lower variability and the lower expense ratio for the investment. The smaller the number of assets needed to mimic is, the smaller the incurred transaction costs will be. Nevertheless, the tracking error is likely to be higher when a small number of assets is used.

To create a tracking portfolio, the simplest technique, called full replication, is to buy appropriate amounts of all the assets that make up the index. Provided that the true index construction weights are available, it allows a perfect tracking. However, it has several disadvantages, one related to the fact that a portfolio can consist of thousands of stocks and the other to the fact that there can be many small or illiquid stocks. These last types of shares increase the risk associated with their sale, which is more difficult, and generate an arbitrage cost that is all the more significant as it is frequent. One of the ways to overcome these drawbacks is to construct a sparse index tracking portfolio (Beasley et al., 2003, Jansen and Van Dijk, 2002) by limiting the number of assets to approximately replicate an index. It corresponds to tracking a signal using a sparse mixture of a given set of time series, see e.g. Benidis et al., 2018. A sparse portfolio simplifies the execution of the portfolio and tends to avoid illiquid stocks that usually correspond to the assets with small weights in an index, since in a sparse setting most of these assets are discarded. Furthermore, since only a small number of assets is used, the transaction costs are reduced significantly due to the reduction of the fixed (minimum) costs in the commission fees. For more details, see (Benidis et al., 2018, Calafiore and El Ghaoui, 2014).

Formulation as a Sparse Regression Problem. Following Calafiore and El Ghaoui, 2014, we give the main steps leading to formulate the Index Tracking problem as Sparse Regression Problem. Let a single financial asset j on which we invest a sum S_j at the beginning of a period. If the rate of return (or return) of this single asset is denoted r_j, we will earn $S_{j,end} = (1 + r_j)S_j$ at the end of the period with $r_j = \dfrac{S_{j,end} - S_j}{S_j}$. For n assets, we define a vector $r \in \mathbb{R}^n$ where the j-th component is the rate of return of the j-th asset. $r(k) \in \mathbb{R}^n$ represents the vector of simple returns of the components

assets during the k-th period of time $[(k-1)\Delta, k\Delta]$, where Δ is a fixed duration.

Let the entries of $x \in \mathbb{R}^n$ are the fractions of an investor's total wealth invested in each of n different assets. Investing at the beginning of the period a total sum S over all assets is made by allocating a fraction $x_j,\ j = 1, \ldots, n$ of S in the j-th asset. The non-negative vector $x \in \mathbb{R}^n_+$ represents the portfolio "mix", and its components sum to one. At the end of the period, the total value of the portfolio is $S_{end} = \sum_{j=1}^{n}(1+r_j)x_j S$. The rate of return of the portfolio is the relative increase in wealth $\frac{S_{end}-S}{S} = \sum_{j=1}^{n}(1+r_j)x_j - 1 = \sum_{j=1}^{n} x_j - 1 + \sum_{j=1}^{n} r_j x_j = r^T x$; i.e., the standard inner product between the vector r of individual returns $r_j,\ j = 1, \ldots, n$ and the vector of the portfolio allocation weights x. The $m \times n$ matrix R gives the (close price) data of the component assets. The component y_k of the vector $y \in \mathbb{R}^T$ represents the return of some target financial index over the j-th period, for $j = 1, \ldots, n$. Vector y is the close price of the target index. Then , the so-called *index tracking* problem is to construct a portfolio x so as to track as close as possible the "benchmark" index returns y. Since the vector of portfolio returns over the considered time horizon is :

$$z = Rx, \quad R \in \mathbb{R}^{m \times n} \quad .$$

We may seek for the portfolio x with minimum Least Squares tracking error, by minimizing $\|Rx - y\|_2^2$. However, we need to take into account the fact that the elements of x represent relative *weights*, that is they are non-negative and they sum up to one. In addition, a cardinality constraint is added for constructing a sparse index tracking portfolio. For given $R \in \mathbb{R}^{m \times n}$ and $y \in \mathbb{R}^m$, this leads to the following sparse regression problem :

$$(P_k) \quad \min_{x \in \mathbb{R}^n_+} \left\{ \|y - Rx\|^2 \;\middle|\; \sum_{i=1}^{n} x_i = 1,\ \mathrm{card}(x) \leq k \right\} \quad .$$

Problem (P_n) is then the problem *without sparsity requirement.* The constraint $x \geq 0, 1^T x = 1$ makes the use of LASSO penalty (constant over the feasible set) irrelevant.

Numerical results. We conducted two experiments with data from Calafiore, 2021. The results have been obtained on a laptop i7-1065G7 CPU@1.30GHz.

In the first experiment, we consider the following sparse techniques with a limited index tracking data set index with $n = 50$ assets over a period of $m = 229$ time steps (the limited number of assets being the limiting dimension that the SDP method can accept) :

(i) *Greedy heuristic*: solve (P_n), take the k greatest value of x and renormalize

(ii) *Reversed greedy heuristic*:

Algorithm 1 Reversed greedy heuristic

```
x ← Solution of (P_n)
while card(x) > k do
    i ← arg min_{1≤j≤n} x_j
    Add the constraint x_i = 0 to (P_n)
    x ← Solution of (P_n)
return x
```

(iii) *SDP approach*: computation of method `sdp2` of Atamturk and Gomez, 2019

(iv) *Mixed-integer programming*: exact solving using `CPLEX`

(v) *Entropy lower bound*: solve the problem $\min_{x \in \mathbb{R}^n_+} \left\{ \|y - Rx\|^2 \;\middle|\; \sum_{i=1}^n x_i = 1,\ B_1(x) \leq k \right\}$

Remark 4.1. *The plain method based on $\|x\|_\infty \geq 1/k$ Pilanci et al., 2012 has also been tested, but it does not produce solutions with significant sparsity for this specific problem.*

We aim at finding a vector $x \in \mathbb{R}^n$ with sparsity $k = 10$. Figures 3 to 5 illustrate the obtained results. For the different methods tested, we carried out various simulations by varying the desired cardinality along the x-axis in an interval ranging from 5 (a high degree of sparsity is required with only 5 non-zero values out of the 50) to 45 (the desired vector is practically dense). The quality of the solution to the problem can be assessed according to two criteria: the value of the objective function at the optimum and the respect of the cardinality constraint. The main comments that can be made

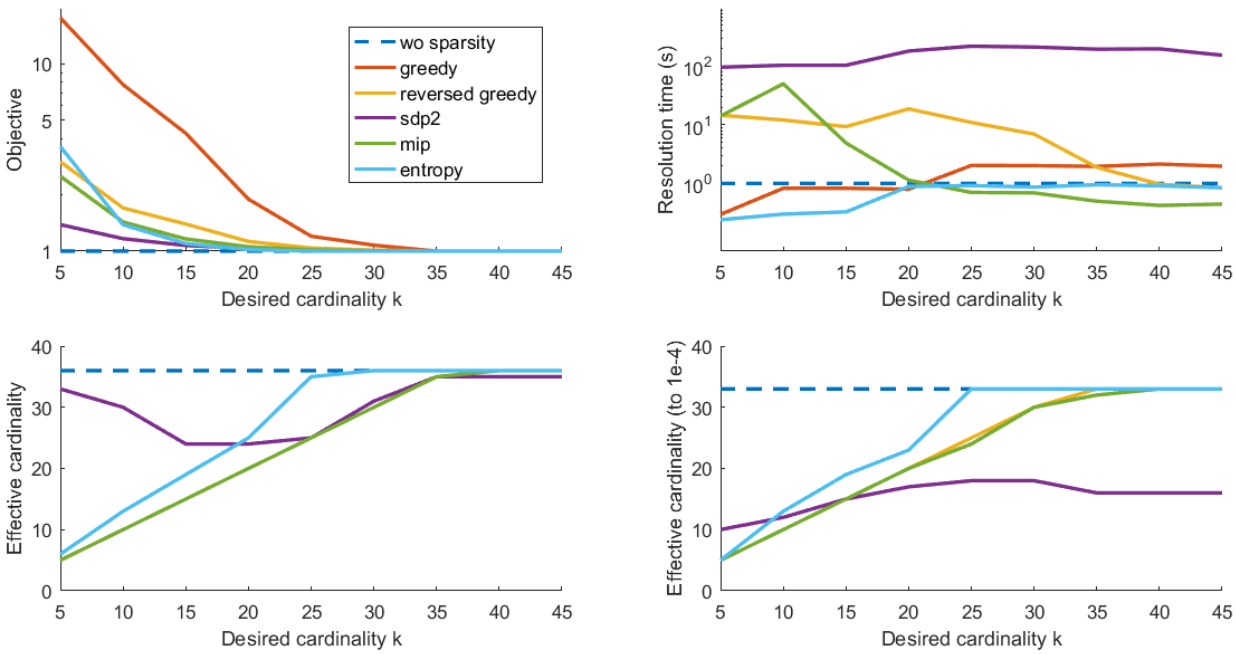


Figure 3: Index tracking for $n = 50$.
At the bottom left, the cardinality constraint is satisfied below the diagonal $y = x$.
At the bottom right, we display the cardinality for the solution filtered with a threshold of 10^{-4}.

from these results are listed below:

(i) Compliance with the cardinality constraint is all the more difficult to satisfy when the desired cardinality is low (see e.g. the time of the exact solver). Beyond a certain degree of sparsity (here, about 30), the problem becomes easy to solve for all the methods tested.

(ii) Concerning the value of the objective function, we note that the "greedy" method is the least efficient of all, while the "reversed greedy" method is competitive. From a desired cardinality of 10, the results of the entropic method are very close to the MIP method (exact resolution of the problem).

(iii) Regarding the respect of the cardinality constraint, we observe that, for strong sparsity requirements, the SDP2 technique absolutely does not respect the desired cardinality unlike the entropic

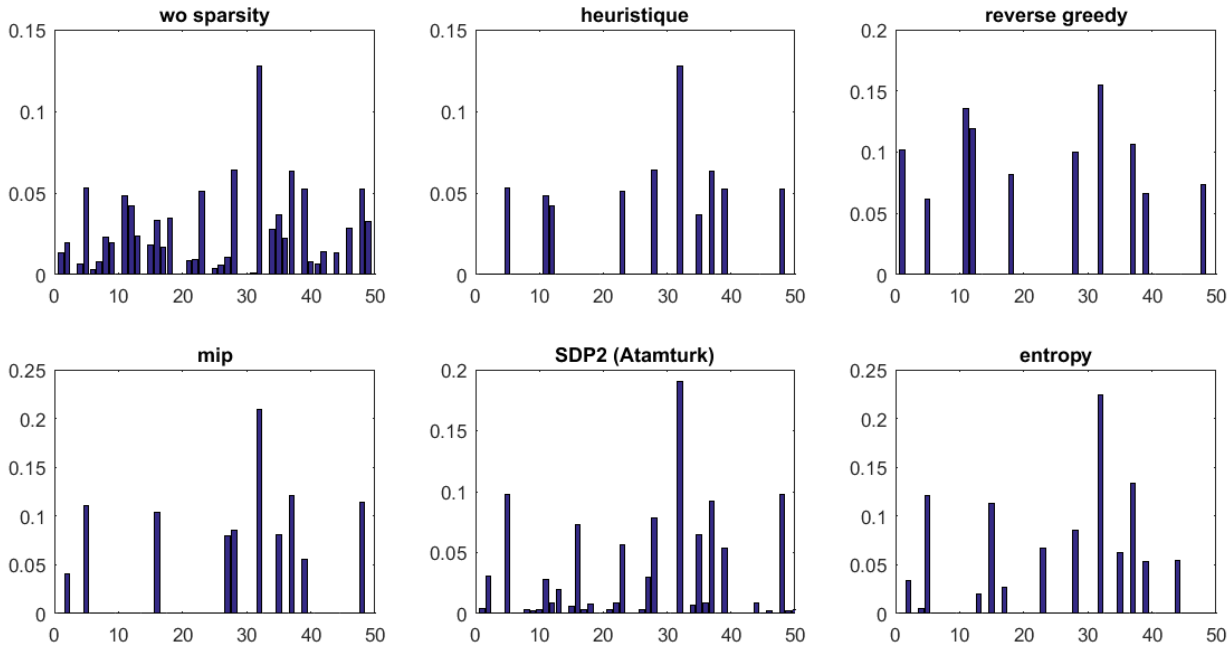


Figure 4: Index tracking for $n = 50$ and $k = 10$

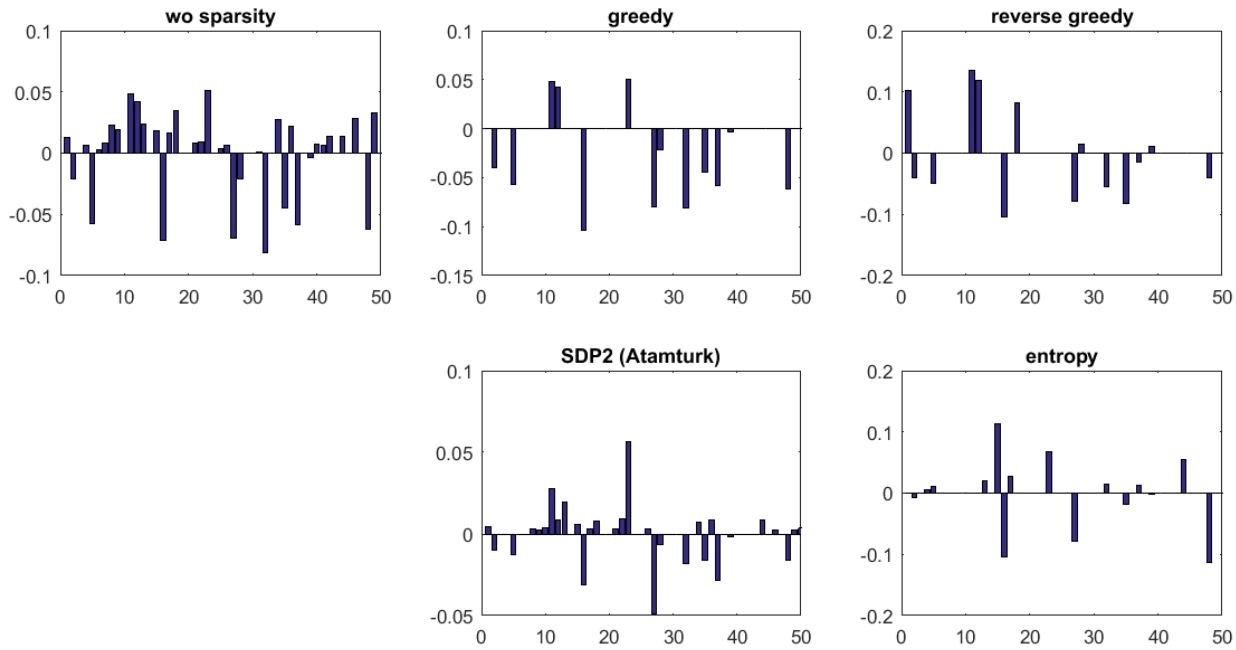


Figure 5: Index tracking for $n = 50$ and $k = 10$.
The bars represent the difference with exact solution.

method. The fact of filtering at 10^{-4} the values of cardinalities obtained does not change the fact that the SDP2 method cannot calculate a solution which respects the desired cardinality. The reversed greedy method provides solutions that respect the cardinality constraint.

(iv) Concerning computation times, unsurprisingly the SDP2 method is the most expensive by far, even for very low sparsity requirements. The exact MIP method is also expensive but the computation time becomes logically lower as the sparsity requirement weakens. The "reversed greedy" method requires a computation time that remains quite high, regardless of the level of cardinality requirement. The entropic method, on the other hand, makes it possible to

calculate solutions in short times, even when the cardinality requirement is very strong. This is an important point in practice, especially for large problem instances, in which the entropic approach is more adapted than the SDP approach or the exact approach.

In a second experiment, we illustrate the possibility to compute a sparse solution via entropic bound even in high dimension ($n = 430$ assets) and hard cardinality requirement ($k = 6$). Figure 6 compares the solution obtained without cardinality constraint (left subfigure) with the relaxed problem (right subfigure). Our technique is highly scalable since its computational time is low (around 1 second for our technique against around 3 seconds for the problem without sparsity requirement). Moreover, the sparsity requirement is almost fully satisfied, as the effective cardinality of the solution is 7 (the target was $k = 6$).

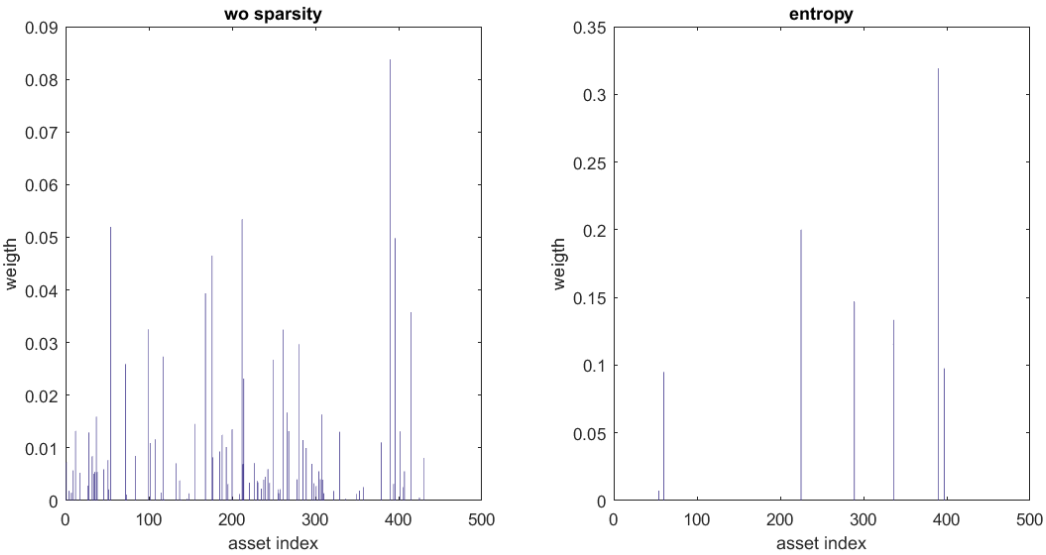


Figure 6: Index Tracking for $n = 430$ and $k = 6$.

5 Conclusions and Perspectives

By using ratios of norms, we proposed a new lower bound of cardinality, based on Shannon entropy. Despite its non-convexity, the use of this entropic bound in a sparse optimization problem is easy, and a local solution can be found very rapidly by using nonlinear solvers. Early results obtained on Index Tracking Finance problem are good regarding other approaches (heuristics, SDP,...) and the proposed approach seems promising.

Among the various perspectives opened to future investigation, we can mention the search for efficient bounds and estimates of cardinality (results on estimates can be found in Bialecki et al., 2015). Extensive simulations on various applications, including Machine Learning, in order to evaluate the efficiency of our approach would be worth considering. Finally, a close look on the relations between Shannon entropy and ℓ_0-pseudonorm should also be done to possibly get approximation guarantees in the sparse optimization problem.

Acknowledgments

We are particularly grateful to G.C. Calafiore who kindly provides the data for the Finance Index Tracking problem.

References

Abboud, Amir, Greg Bodwin, and Seth Pettie (2016). *A Hierarchy of Lower Bounds for Sublinear Additive Spanners*. DOI: 10.48550/ARXIV.1607.07497.

Atamturk, A. and A. Gomez (2019). *Rank-one Convexification for Sparse Regression*. DOI: 10.48550/ARXIV.1901.10334.

Beasley, J.E., N. Meade, and T.J. Chang (2003). "An evolutionary heuristic for the index tracking problem". In: *European Journal of Operational Research* 148.3, pp. 621–643.

Ben Mhenni, R., S. Bourguignon, and J. Ninin (2021). "Global optimization for sparse solution of least squares problems". In: *Optimization Methods and Software*, pp. 1–30. DOI: 10.1080/10556788.2021.1977809.

Benidis, K., Y. Feng, and D. P. Palomar (2018). "Sparse Portfolios for High-Dimensional Financial Index Tracking". In: *IEEE Transactions on Signal Processing* 66.1, pp. 155–170. DOI: 10.1109/tsp.2017.2762286.

Bertsimas, D. and R. Shioda (Nov. 2007). "Algorithm for cardinality-constrained quadratic optimization". In: *Computational Optimization and Applications* 43.1, pp. 1–22. DOI: 10.1007/s10589-007-9126-9.

Bi, J. et al. (Mar. 2003). "Dimensionality Reduction via Sparse Support Vector Machines." In: *Journal of Machine Learning Research* 3, pp. 1229–1243. DOI: 10.1162/153244303322753643.

Bialecki, A., R. Zorgati, and L. El Ghaoui (2014a). "Intra-Day Unit-Commitment : A Group Sparsity Approach". In: *Euro Mini Conference on Stochastic Programming and Energy Applications, Institut Henri Poincaré (IHP), Paris, September 24-26.*

— (2014b). "Intra-Day Unit-Commitment : A Group Sparsity Approach". In: *COPI'14 : Conference on Optimization and Practices in Industry, Palaiseau, October 28-31.*

— (2015). "Estimating the cardinality of a vector for optimization problems". In: *SIAM Conference on Applied Linear Algebra LA'15, Atlanta, October 26-30.*

Bradley, P.S. and O.L. Mangasarian (1998). "Feature Selection via Concave Minimization and Support Vector Machines". In: *Machine Learning Proceedings of the Fifteenth International Conference(ICML98)*. Morgan Kaufmann, pp. 82–90.

Cachin, Christian (1997). "Entropy measures and unconditional security in cryptography". PhD thesis. ETH Zurich.

Calafiore, G.C. (2021). *Index Tracking Data*. Tech. rep. Personnal Communication: Politchnico Torino.

Calafiore, G.C. and L. El Ghaoui (2014). *Optimization Models*. Control systems and optimization series. Cambridge University Press.

Chancelier, J.-P. and M. De Lara (2019). *Lower Bound Convex Programs for Exact Sparse Optimization*. DOI: 10.48550/ARXIV.1902.04813.

Chen, X., F. Xu, and Y. Ye (2010). "Lower Bound Theory of Nonzero Entries in Solutions of L2-Lp Minimization". In: *SIAM J. Sci. Comp.* 32.5, pp. 2832–2852.

Dong, Zhengshan and Wenxing Zhu (2018). "Homotopy Methods Based on l_0 -Norm for Compressed Sensing". In: *IEEE Transactions on Neural Networks and Learning Systems* 29.4, pp. 1132–1146. DOI: 10.1109/TNNLS.2017.2658953.

Donoho, David L (2006). "Compressed sensing". In: *IEEE Transactions on information theory* 52.4, pp. 1289–1306.

Fu, W.J. (1998). "Penalized regression: the bridge versus the lasso". In: *J. Comp. Graph. Stat.* 7, pp. 397–416.

Gasso, G., A. Rakotomamonjy, and S. Canu (2009). "Recovering sparse signals with a certain family of nonconvex penalties and DC programming". In: *IEEE Trans. Sign. Proc.* 57.12, pp. 4686–4698.

Gribonval, R. and M. Nielsen (2003). "Sparse representation in union of bases." In: *IEEE Trans. on Information Theory* 49, pp. 3320–3325.

Guan, W. and A. Gray (2013). "Sparse high-dimensional fractional-norm support vector machine via DC programming". In: *Computational Statistics and Data Analysis* 67, pp. 136–148.

Haddou, M. and Migot T. (2019). "A smoothing method for sparse optimization over convex sets". In: *Optimization Letters* 14.5, pp. 1053–1069. DOI: 10.1007/s11590-019-01408-x.

Hartley, R.V.L. (1928). "Transmission of Information". In: *Bell System Technical Journal* 7.3, pp. 535–563.

Jansen, R. and R. Van Dijk (2002). "Optimal benchmark tracking with small portfolios". In: *The Journal of Portfolio Management* 28.2, pp. 33–39.

Knight, K. and W. Fu (2000). "Asymptotics for lasso-type estimators". In: *Ann. Stat.* 28, pp. 1356–1378.

Koga, Hiroki (2013). "Characterization of the smooth Rényi Entropy Using Majorization". In: *2013 IEEE Information Theory Workshop (ITW)*, pp. 1–5. DOI: 10.1109/ITW.2013.6691332.

Mallat, S. and Z. Zhang (1993). "Matching Pursuit in a Time-Frequency Dictionary". In: *IEEE Trans. Signal Processing* 41.12, pp. 3397–3415.

Marshall, Albert W., Ingram Olkin, and Barry C. Arnold (2011). *Inequalities: Theory of Majorization and Its Applications*. Springer New York. DOI: 10.1007/978-0-387-68276-1.

Nadisic, N. et al. (May 2020). "Exact Sparse Nonnegative Least Squares". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. DOI: 10.1109/icassp40776.2020.9053295.

Ong, C.S. and H.A. Le Thi (Aug. 2013). "Learning sparse classifiers with difference of convex functions algorithms". In: *Optimization Methods and Software* 28.4, pp. 830–854. DOI: 10.1080/10556788.2011.652630.

Pati, Y.C., R. Rezaifar, and P.S. Krishnaprasa (1993). "Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition". In: *27th Asilomar Conf. on Signals Systems and Computers*, pp. 40–44.

Pham Dinh, T. and H.A. Le Thi (2014). "Recent Advances in DC Programming and DCA". In: *Transactions on Computational Collective Intelligence* 8342, pp. 1–37.

Pilanci, M., L. El Ghaoui, and V. Chandrasekaran (2012). "Recovery of Sparse Probability Measures via Convex Programming". In: *Proceedings Conference on Neural Information Processing Systems*.

Rényi, Alfréd et al. (1961). "On measures of entropy and information". In: *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 547-561. Berkeley, California, USA.

Sakai, Y. and K. Iwata (2016). "Extremal relations between shannon entropy and ℓ_α-norm". In: *2016 International Symposium on Information Theory and Its Applications (ISITA)*, pp. 428–432.

Sason, Igal (Nov. 2018). "Tight Bounds on the Rényi Entropy via Majorization with Applications to Guessing and Compression". In: *Entropy* 20.12, p. 896. DOI: 10.3390/e20120896.

Shannon, Claude Elwood (1948). "A mathematical theory of communication". In: *The Bell system technical journal* 27.3, pp. 379–423.

Soubies, E., L. Blanc-Féraud, and G. Aubert (2015). "A Continuous Exact l0 Penalty (CEL0) for Least Squares Regularized Problem". In: *SIAM Journal on Imaging Sciences* 8.3, pp. 1607–1639. DOI: 10.1137/151003714.

Soussen, C. et al. (2011). "From Bernoulli–Gaussian Deconvolution to Sparse Signal Restoration". In: *IEEE Transactions on Signal Processing* 59.10, pp. 4572–4584. DOI: 10.1109/tsp.2011.2160633.

Thiao, M., T. Pham Dinh, and H.A. Le Thi (2008). "DC Programming Approach for a Class of Nonconvex Programs Involving L0 Norm". In: *Modelling, Computation and Optimization in Information Systems and Management Sciences: Second International Conference MCO 2008, Metz, France - Luxembourg, September 8-10, 2008. Proceedings*. Springer Berlin Heidelberg, pp. 348–357.

Tibshirani, R. (1996). "Regression Shrinkage and Selection Via the Lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

Wächter, A. and L.T. Biegler (2006). "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming". In: *Mathematical programming* 106.1, pp. 25–57.

Weston, J. et al. (2003). "Use of the Zero-Norm with Linear Models and Kernel Methods". In: *Journal of Machine Learning Research* 3, pp. 1439–1461.

Zou, H. (2006). "The adaptive lasso and its oracle properties". In: *J. Amer. Stat. Ass.* 101, pp. 1418–1429.

Zou, H. and R. Li (2008). "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models". In: *Ann. Statist.* 36.4, pp. 1509–1533.

A Homotopy

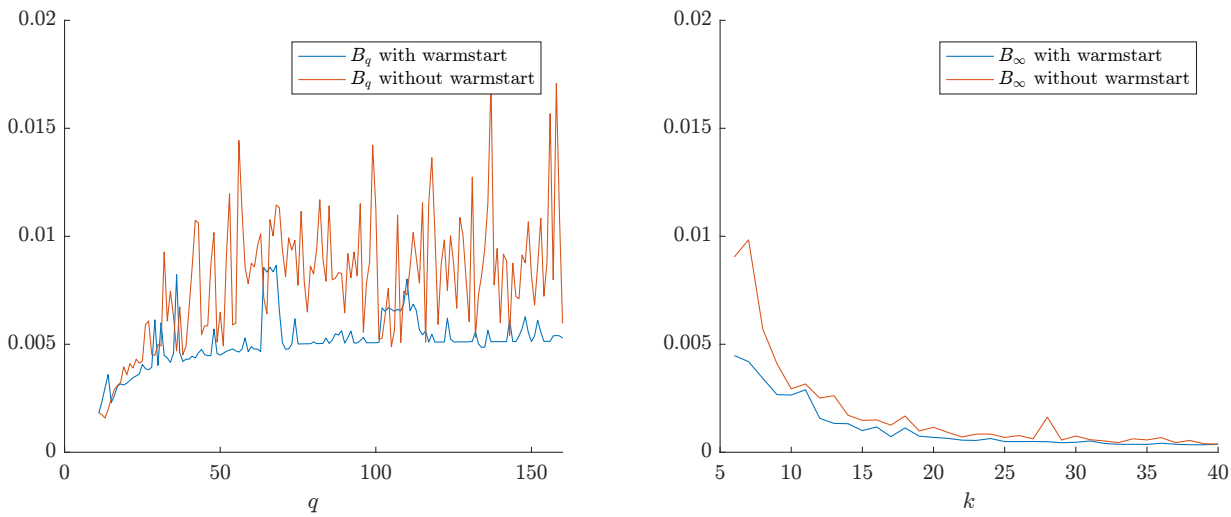


Figure 7: Homotopie in two directions.
In the left subfigure, we plot the solution obtained with the constraint $B_q(x) \leq k$, $k = 6$, starting from $q = 1$ up to $q = 160$.
In the right subfigure, we plot the solution obtained with the constraint $B_\infty(x) \leq k$ starting from $k = 50$ up to $k = 6$

For a reference on homotopy for sparse least-square, see (Abboud et al., 2016) and Dong and Zhu, 2018.