Math. Program., Ser. B (2015) 151:63–87
DOI 10.1007/s10107-015-0894-1

FULL LENGTH PAPER

Sparse learning via Boolean relaxations

Mert Pilanci[1] · **Martin J. Wainwright**[2] · **Laurent El Ghaoui**[1]

Received: 4 November 2014 / Accepted: 17 February 2015 / Published online: 31 March 2015

Abstract We introduce novel relaxations for cardinality-constrained learning problems, including least-squares regression as a special but important case. Our approach is based on reformulating a cardinality-constrained problem exactly as a Boolean program, to which standard convex relaxations such as the Lasserre and Sherali-Adams hierarchies can be applied. We analyze the first-order relaxation in detail, deriving necessary and sufficient conditions for exactness in a unified manner. In the special case of least-squares regression, we show that these conditions are satisfied with high probability for random ensembles satisfying suitable incoherence conditions, similar to results on ℓ_1-relaxations. In contrast to known methods, our relaxations yield lower bounds on the objective, and it can be verified whether or not the relaxation is exact. If it is not, we show that randomization based on the relaxed solution offers a principled way to generate provably good feasible solutions. This property enables us to obtain high quality estimates even if incoherence conditions are not met, as might be expected in real datasets. We numerically illustrate the performance of the relaxation-randomization strategy in both synthetic and real high-dimensional datasets, revealing substantial improvements relative to ℓ_1-based methods and greedy selection heuristics.

✉ Laurent El Ghaoui
elghaoui@berkeley.edu

Mert Pilanci
mert@berkeley.edu

Martin J. Wainwright
wainwrig@berkeley.edu

[1] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA

[2] Department of Electrical Engineering and Computer Sciences and Department of Statistics, University of California, Berkeley, CA, USA

Keywords Sparsity · Regularization · Convex relaxation · Combinatorial optimization · Machine learning

Mathematics Subject Classification 90C25 Convex programming · 90C06 Large-scale problems · 90C09 Boolean programming · 68T05 Learning and adaptive systems

1 Introduction

Over the past several decades, the rapid increase of data dimensionality and complexity has led a tremendous surge of interest of models for high-dimensional data that incorporate some type of low-dimensional structure. Sparsity is a canonical way of imposing low-dimensional structure, and has received considerable attention in many fields, including statistics, signal processing, machine learning and applied mathematics [12,30,34]. Sparse models often typically more interpretable from the scientific standpoint, and they are also desirable from a computational perspective.

The most direct approach to enforcing sparsity in a learning problem is by controlling the ℓ_0-"norm" of the solution, which counts the number of non-zero entries in a vector. Unfortunately, at least in general, optimization problems involving such an ℓ_0-constraint are known to be computationally intractable. The classical approach of circumventing this difficulty while still promoting sparsity in the solution is to replace the ℓ_0-constraint with an ℓ_1-constraint, or alternatively to augment the objective function with an ℓ_1-penalty. This approach is well-known and analyzed under various assumptions on the data generating mechanisms (e.g., [5,6,12,34]). However, in a typical statistical setting, these mechanisms are not under the user's control, and it is difficult to verify post hoc that an ℓ_1-based solution is of suitably high quality.

The main contribution of this paper is to provide a novel framework for obtaining approximate solutions to cardinality-constrained problems, and one in which the quality can be easily verified. Our approach is based on showing a broad class of cardinality-constrained (or penalized) problems can be expressed equivalently as convex programs involving Boolean variables. This reformulation allows us to apply various standard hierarchies of relaxations for Boolean programs, among them Sherali-Adams or Lasserre hierarchies [16,17,29,35]. When the solution of any such relaxation is integral—i.e., belongs to the Boolean hypercube—then it must be an optimal solution to the original problem. Otherwise, any non-integral solution still provides a lower bound on the minimum over all Boolean solutions.

The simplest relaxation is the first-order one, based on relaxing each Boolean variable to the unit interval [0, 1]. We provide an in-depth analysis of the necessary and sufficient conditions for this first-order relaxation to have an integral solution. In the case of least-squares regression, and for a random ensemble of problems of the compressed sensing type [6,12], we show that the relaxed solution is integral with high probability once the sample size exceeds a critical threshold. In this regime, like ℓ_1-relaxations, our first-order method recovers the support of sparse vector exactly, but *unlike* ℓ_1*-relaxations*, the integral solution also certifies that it has recovered the

sparest solution. Finally, there are many settings in which the first-order relaxation might not be integral. For such cases, we study a form of randomized rounding for generating feasible solutions, and we prove a result that controls the approximation ratio. Our framework also allows to specify a target cardinality unlike methods based on ℓ_1 regularization. This feature is desirable for many applications including portfolio optimization [20], machine learning [11,27] and control theory [4].

The remainder of this paper is organized as follows. We begin in Sect. 2 by introducing the problem of sparse learning, and then showing how the constrained version can be reformulated as a convex program in Boolean variables. In Sect. 3, we study the first-order relaxation in some detail, including conditions for exactness as well as analysis of randomized rounding procedures. Section 4 is devoted to discuss of the penalized form of sparse learning problems, whereas Sect. 5 discusses numerical issues and applications to real-world data sets. We conclude the main body with a discussion in Sect. 6, with the majority of our proofs deferred to the Appendix.

2 General sparse learning as a Boolean problem

We consider a learning problem based on samples of the form $(x, y) \in \mathbb{R}^d \times \mathcal{Y}$. This set-up is flexible enough to model various problems, including regression problems (output space $\mathcal{Y} = \mathbb{R}$), binary classification problems (output space $\mathcal{Y} = \{-1, +1\}$), and so on. Given a collection of n samples $\{(x_i, y_i)\}_{i=1}^n$, our goal is to learn a linear function $x \mapsto \langle x, w\rangle$ that can be used to predict or classify future (unseen) outputs. In order to learn the weight vector $w \in \mathbb{R}^d$, we consider a cardinality-constrained program of the form

$$P^* := \min_{\substack{w \in \mathbb{R}^d \\ \|w\|_0 \leq k}} \underbrace{\left\{ \sum_{i=1}^n f(\langle x_i, w\rangle; y_i) + \frac{1}{2}\rho\|w\|_2^2 \right\}}_{F(w)} \tag{1}$$

As will be clarified, the additional regularization term $\frac{1}{2}\rho\|w\|_2^2$ is useful for convex-analytic reasons, in particular in ensuring strong convexity and coercivity of the objective, and thereby the existence of a unique optimal solution $w^* \in \mathbb{R}^d$. Our results also involve the Legendre-Fenchel conjugate of the function $t \mapsto f(t; y)$, given by (for each fixed $y \in \mathcal{Y}$)

$$f^*(s; y) := \sup_{t \in \mathbb{R}} \big\{ s\, t - f(t; y) \big\}. \tag{2}$$

Let us consider some examples to illustrate.

Example 1 (Least-squares regression) In the problem of least-squares regression, the outputs are real-valued (see e.g., [4]). Adopting the cost function $f(t, y) = \frac{1}{2}\big(t - y\big)^2$ leads to ℓ_0-constrained problem

$$P^* := \min_{\substack{w \in \mathbb{R}^d \\ \|w\|_0 \leq k}} \underbrace{\left\{ \frac{1}{2} \sum_{i=1}^n \big(\langle x_i, w\rangle - y_i\big)^2 + \frac{1}{2}\rho\|w\|_2^2 \right\}}_{F_{\mathrm{LS}}(w)} \tag{3}$$

This formulation, while close in spirit to elastic net [39], is based on imposing the cardinality constraint exactly, as opposed to in a relaxed form via ℓ_1-regularization. However, in contrast to the elastic net, it is a nonconvex problem, so that we need to study relaxations of it. A straightforward calculation yields the conjugate dual function

$$f^*(s; y) = \frac{s^2}{2} + s\, y, \tag{4}$$

which will play a role in our relaxations of the nonconvex problem (3).

The preceding example has a natural extension in terms of generalized linear models:

Example 2 (Generalized linear models) In a generalized linear model, the output $y \in \mathcal{Y}$ is related to the covariate $x \in \mathbb{R}^d$ via a conditional distribution in the exponential form (see e.g. [21,24])

$$\mathbb{P}_w(y \mid x) = h(y) \exp\big(y \langle x,\ w\rangle - \psi(\langle x,\ w\rangle)\big). \tag{5}$$

Here $h : \mathbb{R}^d \to \mathbb{R}_+$ is some fixed function, and $\psi : \mathbb{R} \to \mathbb{R}$ is the cumulant generating function, given by $\psi(t) = \log \int_{\mathcal{Y}} e^{ty} h(y) dy$. Letting $f(\langle x,\ w\rangle; y)$ be the negative log-likelihood associated with this family, we obtain the general family of cardinality-constrained likelihood estimates

$$\min_{\substack{w \in \mathbb{R}^d \\ \|w\|_0 \le k}} \underbrace{\Big\{ \sum_{i=1}^n \big\{\psi(\langle x_i,\ w\rangle\big) - y_i \langle x_i,\ w\rangle\big\} + \frac{1}{2}\rho \|w\|_2^2 \Big\}}_{F_{\mathrm{GR}}(w)} \tag{6}$$

Specifically, least-squares regression is a particular case of the problem (6), corresponding to the choice $\psi(t) = t^2/2$. Similarly, logistic regression for binary responses $y \in \{0, 1\}$ can be obtained by setting $\psi(t) = \log(1 + e^t)$.

In the likelihood formulation (6), we have $f(t; y) = \psi(t) - yt$, whence conjugate dual takes the form

$$f^*(s; y) = \sup_{t \in \mathbb{R}} \big\{st - \psi(t) + yt\big\} \ = \ \psi^*(s + y), \tag{7}$$

where ψ^* denotes the conjugate dual of ψ. As particular examples, in the case of logistic regression, the dual of the logistic function $\psi(t) = \log(1 + e^t)$ takes the form $\psi^*(s) = s \log s + (1 - s) \log(1 - s)$ for $s \in [0, 1]$, and takes the value infinity otherwise.

As a final example, let us consider a cardinality-constrained version of the support vector machine:

Example 3 (Support vector machine classification) In this case, the outputs are binary $y \in \{-1, 1\}$, and our goal is to learn a linear classifier $x \mapsto \mathrm{sign}(\langle x,\ w\rangle) \in \{-1, 1\}$ [7].

The cardinality-constrained version of the support vector machine (SVM) is based on minimizing the objective function

$$\min_{\substack{w\in\mathbb{R}^d \\ \|w\|_0\le k}} \underbrace{\Big\{\sum_{i=1}^{n}\phi\big(y_i\,\langle x_i,\,w\rangle\big)+\frac{1}{2}\rho\|w\|_2^2\Big\}}_{F_{\mathrm{SVM}}(w)}, \tag{8}$$

where $\phi(t)=\max\{1-t,0\}$ is known as the hinge loss function. The conjugate dual of the hinge loss takes the form

$$\phi^*(s)=\begin{cases} s & \text{if } s\in[-1,0] \\ \infty & \text{otherwise.}\end{cases}$$

Having considered various examples of sparse learning, we now turn to developing an exact Boolean representation that is amenable to various relaxations.

2.1 Exact representation as a Boolean convex program

Let us now show how the cardinality-constrained program (1) can be represented exactly as a convex program in Boolean variables. This representation, while still nonconvex, is useful because it immediately leads to a hierarchy of relaxations. Given the collection of covariates $\{x_i\}_{i=1}^n$, we let $X\in\mathbb{R}^{n\times d}$ denote the design matrix with $x_i^T\in\mathbb{R}^d$ as its ith row.

Theorem 1 (Exact representation) *Suppose that for each $y\in\mathcal{Y}$, the function $t\mapsto f(t;y)$ is closed and convex. Then for any $\rho>0$, the cardinality-constrained program (1) can be represented exactly as the Boolean convex program*

$$P^*=\min_{\substack{u\in\{0,1\}^d \\ \sum_{j=1}^d u_j\le k}} \underbrace{\max_{v\in\mathbb{R}^n}\Big\{-\frac{1}{2\rho}v^TXD(u)X^Tv-\sum_{i=1}^{n}f^*(v_i;y_i)\Big\}}_{G(u)}, \tag{9}$$

where $D(u):=\mathrm{diag}(u)\in\mathbb{R}^{d\times d}$ is a diagonal matrix.

The function $u\mapsto G(u)$—in particular, defined by maximizing over $v\in\mathbb{R}^n$—is a maximum of a family of functions that are linear in the vector u, and hence is convex. Thus, apart from the Boolean constraint, all other quantities in the program (9) are relatively simple: a linear constraint and a convex objective function. Consequently, we can obtain tractable approximations by relaxing the Boolean constraint. The simplest such approach is to replace the Boolean hypercube $\{0,1\}^d$ with the unit hypercube $[0,1]^d$. Doing so leads the *interval relaxation* of the exact Boolean representation, namely the convex relaxation

$$P_{\text{IR}} = \min_{\substack{u\in[0,1]^d \\ \sum_{j=1}^d u_j \le k}} \underbrace{\max_{v\in\mathbb{R}^n} \Big\{ -\frac{1}{2\rho} v^T X D(u) X^T v - \sum_{i=1}^n f^*(v_i; y_i) \Big\}}_{G(u)}. \tag{10}$$

Note that this is a convex program, and so can be solved by standard methods. In particular the sub-gradient descent method (e.g., see [25]) can be applied directly if a closed form solution, or a solver for the inner maximization problem is available. In Sect. 3, we return to analyze when the interval relaxation is tight—that is, when $P_{\text{IR}} = P^*$.

In the case of least-squares regression, Theorem 1 and the interval relaxation take an especially simple form, which we state as a corollary.

Corollary 1 *The cardinality constrained problem is equivalent to the Boolean SDP*

$$P^* = \min_{\substack{(u,t)\in\{0,1\}^d\times\mathbb{R}_+ \\ \sum_{j=1}^d u_j \le k}} t \quad \text{such that} \begin{bmatrix} I_n + \frac{1}{\rho} X D(u) X^T & y \\ y^T & t \end{bmatrix} \succeq 0. \tag{11}$$

Thus, the interval relaxation (10) *is an ordinary SDP in variables* $(u, t) \in [0, 1]^d \times \mathbb{R}_+$.

Proof As discussed in Example 1, the conjugate dual of the least-squares loss $t \mapsto f(t; y) = \frac{1}{2}(t - y)^2$ is given by $f^*(s; y) = \frac{s^2}{2} + sy$. Substituting this dual function into Eq. (9), we find that

$$G(u) = \max_{v\in\mathbb{R}^n} \Big\{ -\frac{1}{2} v^T \left(\frac{X D(u) X^T}{\rho} + I \right) v - \langle v,\ y \rangle \Big\},$$

where we have defined the diagonal matrix $D(u) := \text{diag}(u) \in \mathbb{R}^{d\times d}$. Taking derivatives shows that the optimum is achieved at

$$\widehat{v} = -\left(\frac{X D(u) X^T}{\rho} + I \right)^{-1} y, \tag{12}$$

and substituting back into Eq. (9) and applying Theorem 1 yield the representation

$$P^* = \min_{\substack{u\in\{0,1\}^d \\ \sum_{j=1}^d u_j \le k}} \Big\{ y^T \left(\frac{1}{\rho} X D(u) X^T + I_n \right)^{-1} y \Big\}. \tag{13}$$

By introducing a slack variable $t \in \mathbb{R}_+$ and using the Schur complement formula (see e.g. [4]), some further calculation shows that this Boolean problem (13) is equivalent to the Boolean SDP (11), as claimed. □

We now present the proof of Theorem 1.

Proof Recalling that $D(u) := \mathrm{diag}(u)$ is a diagonal matrix, for each fixed $u \in \{0, 1\}^d$, consider the change of variable $w \mapsto D(u)w$. With this notation, the original problem (1) is equivalent to

$$P^* = \min_{\|D(u)w\|_0 \le k} \Big\{ \sum_{i=1}^n f(\langle D(u)x_i,\ w\rangle; y_i) + \frac{1}{2}\rho\|D(u)w\|_2^2 \Big\}. \tag{14}$$

Noting that we can take $w_i = 0$ when $u_i = 0$ and vice-versa, the original problem (1) becomes

$$P^* = \min_{\substack{u\in\{0,1\}^d \\ \sum_{j=1}^d u_j \le k}} \ \min_{w\in\mathbb{R}^d} \Big\{ \sum_{i=1}^n f(\langle D(u)x_i,\ w\rangle; y_i) + \frac{1}{2}\rho\|w\|_2^2 \Big\}. \tag{15}$$

It remains to prove that, for each fixed Boolean vector $u \in \{0, 1\}^d$, we have

$$\begin{aligned} &\min_{w\in\mathbb{R}^d} \Big\{ \sum_{i=1}^n f(\langle D(u)x_i,\ w\rangle; y_i) + \frac{1}{2}\rho\|w\|_2^2 \Big\} \\ &\quad = \max_{v\in\mathbb{R}^n} \Big\{ -\frac{1}{2\rho}\|D(u)X^T v\|_2^2 - \sum_{i=1}^n f^*(v_i; y_i) \Big\}. \end{aligned} \tag{16}$$

From the conjugate representation of f, we find that

$$\min_{w\in\mathbb{R}^d} \max_{v\in\mathbb{R}^n} \Big\{ \sum_{i=1}^n v_i\langle D(u)x_i,\ w\rangle - f^*(v_i; y_i) + \frac{1}{2}\rho\|w\|_2^2 \Big\}.$$

Under the stated assumptions, strong duality must hold, so that it is permissible to exchange the order of the minimum and maximum. Doing so yields

$$\max_{v\in\mathbb{R}^n} \min_{w\in\mathbb{R}^d} \Big\{ \sum_{i=1}^n v_i\langle D(u)x_i,\ w\rangle - f^*(v_i; y_i) + \frac{1}{2}\rho\|w\|_2^2 \Big\}.$$

Finally, strong convexity ensures that the minimum over w is unique: more specifically, it is given by $w^* = \frac{1}{\rho}\sum_{i=1}^n D(u)x_i v_i$. Substituting this optimum yields the claimed equality (16). □

3 Convex-analytic conditions for IR exactness

We now turn to analysis of the interval relaxation (10), and in particular, determining when it is exact. Note that by strong convexity, the original cardinality-constrained problem (1) has a unique solution, say $w^* \in \mathbb{R}^d$. Let S denote the support set of w^*,

and let u^* be a Boolean indicator vector for membership in S—that is, $u_j^* = 1$ if $j \in S$ and zero otherwise.

An attractive feature of the IR relaxation is that integrality of an optimal solution $\widehat{u}$ to the relaxed problem provides a *certificate of exactness*—that is, if the interval relaxation (10) has an optimal solution $\widehat{u} \in \{0, 1\}^d$, then it must be the case that $\widehat{u} = u^*$ (so that we recover the support set of w^*), and moreover that

$$P_{\text{IR}} = P^*. \tag{17}$$

In this case, we are guaranteed to recover the optimal solution w^* of the original problem (1) by solving the constrained problem with $w_j = 0$ for all $j \notin S$.

In contrast, methods based on ℓ_1-relaxations do not provide such certificates of exactness. In the least-squares regression, the use of ℓ_1-relaxation is known as the Lasso [30], and there is an extensive literature devoted to conditions on the design matrix $X \in \mathbb{R}^{n \times d}$ under which the ℓ_1-relaxation provides a "good" solution. Unfortunately, these conditions are either computationally infeasible to check (e.g., restricted eigenvalue, isometry and nullspace conditions [3,9] and the related irrepresentability conditions for support recovery [14,22,38]). Although polynomial-time checkable conditions do exist (such as pairwise incoherence conditions [13,14,31]), they provide weak guarantees, only holding for sample sizes much larger than the threshold at which the ℓ_1-relaxation begins to work. In addition, most of the previous work on analyzing ℓ_1 relaxations considered a statistical data model where there exists a true sparse coefficient generating the response. However in many applications such assumptions do not necessarily hold and it is unclear whether ℓ_1 regularization provides a good optimization heuristic for an arbitrary input data.

It is thus of interest to investigate conditions under which the relaxation (IR) is guaranteed to have an integer solution and hence be tight. The following result provides an if-and-only if characterization.

Proposition 1 *The interval relaxation is tight—that is, $P_{IR} = P^*$—if and only if there exist a pair $(\lambda, \widehat{v}) \in \mathbb{R}_+ \times \mathbb{R}^n$ such that*

$$\widehat{v} \in \arg\max_{v \in \mathbb{R}^n} \Big\{ -\frac{1}{2\rho} v^T X_S X_S^T v - \sum_{i=1}^n f^*(v_i; y_i) \Big\}, \quad \text{and} \tag{18a}$$

$$|\langle X_j, \widehat{v} \rangle| > \lambda \ \text{ for all } j \in S, \text{ and } \quad |\langle X_j, \widehat{v} \rangle| < \lambda \ \text{ for all } j \notin S, \tag{18b}$$

where $X_j \in \mathbb{R}^n$ denotes the jth column of the design matrix, S denotes the support of the unique optimal solution w^* to the original problem (1).

Proof Beginning with the saddle-point representation from Eq. (10), we apply the first-order convex optimality condition for constrained minimization. More precisely, the relaxed solution $\widehat{u}$ is optimal if and only if the following inclusion holds:

$$0 \in \left\{ \partial_u \max_{v \in \mathbb{R}^n} \Big\{ -\frac{1}{2\rho} v^T X D(u) X^T v - \sum_{i=1}^n f^*(v_i; y_i) \Big\} + \mathbb{N} \right\},$$

where $\mathbb{N}$ denotes the normal cone of the constraint set $\left\{u \in [0,1]^d \mid \sum_{j=1}^d u_j \leq k\right\}$. Note that the subgradient with respect to u_j is given by $-(\langle X_j, \widehat{v}\rangle)^2$, where the vector $\widehat{v}$ was defined in Eq. (18a). Using representation of the normal cone at the integral point u^* and associating $\lambda \geq 0$ as the dual parameter corresponding to constraint $\sum_{j=1}^d u_j$, we arrive at the stated condition (18b). □

In the case of least-squares regression, the conditions of Proposition 1 can be simplified substantially. Recall that interval relaxation for least-squares regression is given by

$$P_{IR} = \min_{\substack{u \in [0,1]^d \\ \sum_{j=1}^d u_j \leq k}} \left\{ y^T \left(\frac{1}{\rho} X D(u) X^T + I_n\right)^{-1} y \right\}. \tag{19}$$

Let S denote the support of the unique optimal solution w^* to the original least-squares problem (3), say of cardinality k, and define the $n \times n$ matrix

$$M := \left(I_n + \rho^{-1} X_S X_S^T\right)^{-1} \tag{20}$$

With this notation, we have:

Corollary 2 *The interval relaxation of cardinality-constrained least-squares is exact ($P_{IR} = P^*$) if and only there exists a scalar $\lambda \in \mathbb{R}_+$ such that*

$$\left|X_j^T M y\right| > \lambda \quad \text{for all } j \in S, \text{ and} \tag{21a}$$

$$\left|X_j^T M y\right| \leq \lambda \quad \text{for all } j \notin S, \tag{21b}$$

where $X_j \in \mathbb{R}^n$ denotes the jth column of X.

Proof From the proof of Corollary 1, recall the Boolean convex program (13). As shown in Eq. (12), its optimum is achieved at $\widehat{v} = -(I_n + X D(u^*) X^T) y$, where u^* is a Boolean indicator for membership in S. Applying Proposition 1 with this choice of $\widehat{v}$ yields the necessary and sufficient conditions

$$\left|y^T (\rho I_n + X D(u^*) X^T)^{-1} X_j\right| > \lambda \quad \text{for all } j \in S, \text{ and}$$

$$\left|y^T (\rho I_n + X D(u^*) X^T)^{-1} X_j\right| \leq \lambda \quad \text{for all } j \in S^c,$$

and completes the proof. □

In order to gain an understanding of the above corollary consider an example where the rows of X_S are orthonormal and $n = k$, hence $M = (I_n + \rho^{(} - 1) I_n)^{-1} = \rho/(1+\rho) I_n$. Then the conditions for integrality reduce to checking whether there exists $\lambda' \in \mathbb{R}_+$ such that

$$\left|X_j^T y\right| > \lambda' \quad \text{for all } j \in S, \text{ and}$$

$$\left|X_j^T y\right| \leq \lambda' \quad \text{for all } j \notin S.$$

Intuitively the above condition basically checks if the columns in the correct support are more aligned to the response y compared to the columns outside the support.

Also note that by the matrix inversion formula, we have the alternative representation,

$$M = \left(I_n + \rho^{-1} X_S X_S^T\right)^{-1} = I_n - X_S\left(\rho I_d + X_S^T X_S\right)^{-1} X_S^T \ ,$$

For random ensembles, Corollary 2 allows the use of a primal witness method to certify exactness of the IR method. In particular, if we can construct a scalar λ for which the two bounds (21a) and (21b) hold with high probability, then we can certify exactness of the relaxation. We illustrate this approach in the following subsection.

3.1 Sufficient conditions for random ensembles

In order to assess the performance of the interval relaxation (10), we performed some simple experiments for the least squares case, first generating a design matrix $X \in \mathbb{R}^{n \times d}$ with i.i.d. $N(0, 1)$ entries, and then forming the response vector $y = Xw^* + \epsilon$, where the noise vector $\epsilon \in \mathbb{R}^n$ has i.i.d. $N(0, \gamma^2)$ entries. The unknown regression vector w^* was k-sparse, with absolute entries of the order $1/\sqrt{k}$ on its support. Each such problem can be characterized by the triple (n, d, k) of sample size, dimension and sparsity, and the question of interest is to understand how large the sample size should be in order to ensure exactness of a method. For instance, for this random ensemble, the Lasso is known [33] to perform exact support recovery once $n \gtrsim k \log(d - k)$, and this scaling is information-theoretically optimal [32]. Does the interval relaxation also satisfy this same scaling?

In order to test the IR relaxation, we performed simulations with sample size $n = \alpha k \log d$ for a control parameter $\alpha \in [2, 8]$, for three different problem sizes $d \in \{64, 128, 256\}$ and sparsity $k = \lceil \sqrt{d} \rceil$. Figure 1 shows the probability of successful recovery versus the control parameter α for these different problem sizes, for both the Lasso and the IR method. Note that both methods undergo a phase transition once the sample size n is larger than some constant multiple of $k \log(d - k)$.

The following result provides theoretical justification for the phase transition behavior exhibited in Fig. 1:

Theorem 2 *Suppose that we are given a sample size* $n > c_0 \frac{\gamma^2 + \|w_S^*\|_2^2}{w_{min}^2} \log d$, *and that we solve the interval relaxation with* $\rho = \sqrt{n}$. *Then with probability at least* $1 - 2e^{-c_1 n}$, *the interval relaxation is integral, so that* $P_{IR} = P^*$.

For a typical k-sparse vector, we have $\frac{\|w^*\|_2^2}{w_{\min}^2} \approx k$, so that Theorem 2 predicts that the interval relaxation should succeed with $n \geq k \log(d - k)$ samples, as confirmed by the plots in Fig. 1.

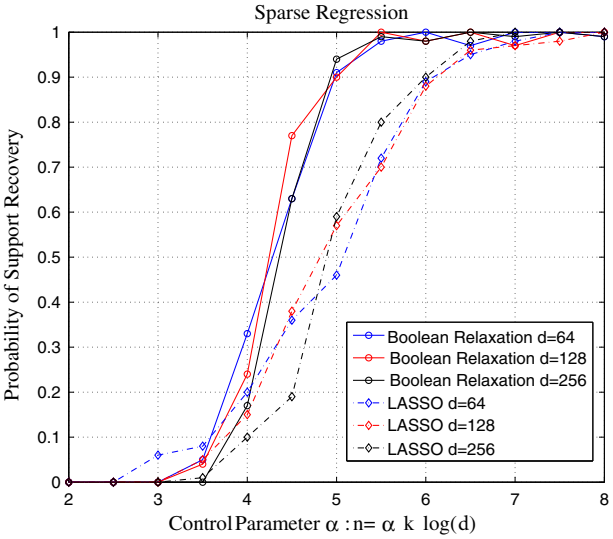

Fig. 1 Problem of exact support recovery for the Lasso and the interval relaxation for different problem sizes $d \in \{64, 128, 256\}$. As predicted by theory, both methods undergo a phase transition from failure to success once the control parameter $\alpha := \frac{n}{k \log(d-k)}$ is sufficiently large. This behavior is confirmed for the interval relaxation in Theorem 2

3.2 Analysis of randomized rounding

In this section, we describe a method to improve the interval relaxation scheme introduced earlier. The convex relaxation of the Boolean hypercube constraint $u \in \{0, 1\}^d$ to the standard hypercube constraint $u \in [0, 1]^d$ might produce an integral solution—in particular, when the conditions in Proposition 1 are not satisfied. In this case, it is natural to consider how to use the fractional solution $\widehat{u} \in [0, 1]^d$ to produce a feasible Boolean solution $\widetilde{u} \in \{0, 1\}^d$. By construction, the objective function values $(G(\widehat{u}), G(\widetilde{u}))$ defined by this pair will sandwich the optimal value—viz

$$G(\widehat{u}) \;\leq\; P^* \;\leq\; G(\widetilde{u}).$$

Here G is the objective function from the original Boolean problem (9).

Randomized rounding is a classical technique for converting fractional solutions into integer solutions with provable approximation guarantees [23]. Here we consider the simplest possible form of randomized rounding in application to our relaxation. Given the fractional solution $\widehat{u} \in [0, 1]^d$, suppose that we generate a feasible Boolean solution $\widetilde{u} \in \{0, 1\}^d$ as follows

$$\mathbb{P}[\widetilde{u}_i = 1] = \widehat{u}_i \quad \text{and} \quad \mathbb{P}[\widetilde{u}_i = 0] = 1 - \widehat{u}_i. \tag{22}$$

By construction, this random Boolean vector matches the fractional solution in expectation—that is, $\mathbb{E}[\widetilde{u}] = \widehat{u}$, and moreover its expected ℓ_0-norm is given by

$$\mathbb{E}[\|\widetilde{u}\|_0] = \sum_{i=1}^{d} \mathbb{P}[\widetilde{u}_i = 1] \; = \; \sum_{i=1}^{d} \widehat{u}_i \leq k,$$

where the final inequality uses the feasibility of the fractional solution $\widehat{u}$. The random Boolean solution $\widetilde{u}$ can be used to define a randomized solution $\tilde{w} \in \mathbb{R}^d$ of the original problem via

$$\tilde{w} = \arg\min_{w \in \mathbb{R}^d} F\big(D(\widetilde{u})w\big), \tag{23}$$

where the function F was defined in Eq. (1).

Without loss of generality, consider the least squares problem and assume the columns are normalized, i.e., $\|x_j\|_2 = 1$ for $j = 1, \ldots, d$ and $\|y\|_2 = 1$, then we have the following result. Let $R \subset \{1, \ldots, d\}$ be the subset of coordinates on which $\widehat{u}$ takes fractional values (i.e., $\widehat{u}_j \in (0, 1)$ for all $j \in R$) and let $r = |R|$ be the cardinality of this set.

Theorem 3 *There are universal constants c_j such that for any $\delta \in (0, 1)$, with probability at least $1 - c_1 e^{-c_2 k\delta^2} - \frac{1}{\min\{r,n\}^{c_3}}$, the randomly rounded solution $\tilde{w}$ has ℓ_0-norm at most $(1 + \delta)k$, and has optimality gap at most*

$$F(\tilde{w}) - P^* \leq c_4 \frac{\sqrt{r \log \min\{r, n\}}}{\rho}. \tag{24}$$

Note that the optimality gap in the preceding bound is negligible when the number of fractional solutions are small enough, and vanishes when the solution is integral, i.e., $r = 0$. The optimality gap also decreases when ρ gets larger in which case the objective of the original problem is heavily regularized by $\frac{\rho}{2}\|w\|_2^2$. The bound in Theorem 3 uses concentration bounds from random matrix theory [1] which are known to be sharp estimates of the statistical deviation in random sampling.

In our simulations, in order to be sure that we compare with a feasible integral solution (i.e., with at most k entries), we generate T realizations—say $\{\widetilde{u}^1, \ldots, \widetilde{u}^T\}$ of the rounding procedure—and then pick the one $\widetilde{u}^*$ that has smallest objective value $G(\widetilde{u})$ among the feasible solutions. (Note that $\widetilde{u}^*$ will exist with high probability for reasonable choices of T.) Finally, we define $\tilde{w}^* = \arg\min_w F(D(\widetilde{u}^*))$. Denoting this procedure as *randomized rounding of order T*, we study its empirical behavior in Sect. 5 in the sequel.

The computational complexity of the randomized rounding procedure is dominated by evaluating $F\big(D(\widetilde{u})w\big)$ a total of T times. However since $\widetilde{u}$ are sparse vectors this procedure is very efficient. For the least squares problem with target cardinality k the complexity becomes $\mathcal{O}(Tk^2n)$ since evaluating $\big(D(\widetilde{u})w\big)$ can be done in $\mathcal{O}(k^2n)$ time using QR decomposition.

We note that in some other applications there might be additional constraints imposed on the vector u such as block sparsity or graphical structure. In such cases the randomized rounding process needs to be altered accordingly, or variants of rejection sampling can be used to generate vectors until constraints are satisfied.

4 Penalized forms of cardinality

Up to this point, we have consider the cardinality-constrained versions of sparse learning problems. If we instead enforce sparsity by augmenting the objective with some multiple of the ℓ_0-norm, this penalized objective can also be reformulated as Boolean program with a convex objective.

4.1 Reformulation as Boolean program

More precisely, suppose that we begin with the cardinality-penalized program

$$P^*(\lambda) := \min_{w \in \mathbb{R}^d} \Big\{ \sum_{i=1}^n f(\langle x_i, w \rangle; y_i) + \frac{1}{2}\rho \|w\|_2^2 + \lambda \|w\|_0 \Big\}. \tag{25}$$

As before, we suppose that for each $y \in \mathcal{Y}$, the function $t \mapsto f(t; y)$ is closed and convex. Under this condition, the following result provides an equivalent formulation as a convex program in Boolean variables:

Theorem 4 *For any $\rho > 0$ and $\lambda > 0$, the cardinality-penalized program (25) can be represented exactly as the Boolean convex program*

$$P^*(\lambda) = \min_{u \in \{0,1\}^d} \max_{v \in \mathbb{R}^n} \Big\{ -\frac{1}{2\rho} v^T X D(u) X^T v - \sum_{i=1}^n f^*(v_i; y_i) + \lambda \sum_{i=1}^d u_i \Big\}, \tag{26}$$

where $D(u) := \mathrm{diag}(u) \in \mathbb{R}^{d \times d}$ is a diagonal matrix.

The proof is very similar to that of Theorem 1, and so we omit it.

As a consequence of the equivalent Boolean form (26), we can also obtain various convex relaxations of the cardinality-penalized program. For instance, the first-order relaxation takes the form

$$P_{\mathrm{IR}}(\lambda) = \min_{u \in [0,1]^d} \max_{v \in \mathbb{R}^n} \Big\{ -\frac{1}{2\rho} v^T X D(u) X^T v - \sum_{i=1}^n f^*(v_i; y_i) + \lambda \sum_{i=1}^d u_i \Big\}, \tag{27}$$

which is the analogue of our first-order relaxation (13) for the constrained version of sparse learning.

As with our previous analysis, it is possible to eliminate the minimization over u from this saddle point expression. Strong duality holds, so that the maximum and minimum may be exchanged. In order to evaluate the minimum over u, we observe that $\frac{1}{2\rho} v^T X D(u) X^T v = \sum_{i=1}^d u_i \big(\frac{1}{2\rho}(x_i^T v)^2\big)$, and moreover that

$$\min_{u \in [0,1]^d} \Big\{ -\sum_{i=1}^d u_i \left(\frac{1}{2\rho}(x_i^T v)^2 - \lambda \right) \Big\} = -\sum_{i=1}^d \left(\frac{1}{2\rho}(x_i^T v)^2 - \lambda \right)_+,$$

Putting together the pieces, we can write the interval relaxation in the penalized case as the following convex (but non-differentiable) program

$$P_{\text{IR}}(\lambda) = \max_{v \in \mathbb{R}^n} \Big\{ -\sum_{i=1}^{d} \Big(\frac{1}{2\rho} (x_i^T v)^2 - \lambda \Big)_+ - \sum_{i=1}^{n} f^*(v_i; y_i) \Big\}. \tag{28}$$

4.2 Least-squares regression

As before, the relaxation (28) takes an especially simple form for the special but important case of least-squares regression. In particular, in the least-squares case, we have $f(t, y) = \frac{1}{2}(t - y)^2$, along with the corresponding conjugate dual function $f^*(s; y) = \frac{s^2}{2} + s\, y$. Consequently, the general relaxation (28) reduces to

$$P_{\text{IR}}(\lambda) = \max_{v \in \mathbb{R}^n} \Big\{ -\sum_{i=1}^{d} \Big(\frac{1}{2\rho} (x_i^T v)^2 - \lambda \Big)_+ - v^T y - \frac{1}{2} \|v\|_2^2 \Big\}, \tag{29}$$

As we now show, this convex program is equivalent to minimizing the least-squares objective using a form of regularization that combines the ℓ_1 and ℓ_2-norms. In particular, let us define

$$B(t) = \frac{1}{2} \min_{z \in [0,1]} \Big\{ z + \frac{t^2}{z} \Big\} = \begin{cases} |t| & \text{if } |t| \leq 1 \\ \frac{t^2+1}{2} & \text{otherwise} \end{cases}. \tag{30}$$

This function combines the ℓ_1 and ℓ_2 norms in the way that is the opposite Huber's robust penalty; consequently, we call it the *reverse Huber penalty*.

Corollary 3 *The interval relaxation* (29) *for the cardinality-penalized least-squares problem has the equivalent form*

$$P_{IR}(\lambda) = \min_{w \in \mathbb{R}^d} \Big\{ \frac{1}{2} \|Xw - y\|_2^2 + 2\lambda \sum_{i=1}^{d} B\Big(\frac{\sqrt{\rho} w_i}{\sqrt{\lambda}} \Big) \Big\}, \tag{31}$$

where B denotes the reverse Huber penalty.

A plot of the reverse Huber penalty is displayed in Fig. 2 and compared with the ℓ_1-norm $t \mapsto \lambda|t|$, as well as the ℓ_0-based penalty $t \mapsto \lambda \|t\|_0 + \frac{1}{2} t^2$.

Proof Consider the representation (29) for the least-squares case. We can represent the coordinatewise functions $(\cdot)_+$ function using a vector $p \in \mathbb{R}^d$ of auxiliary variables as follows

$$\begin{aligned} P_{\text{IR}}(\lambda) = \max_{v,\, p} & \Big\{ -1^T p - \frac{1}{2} \|v\|_2^2 - \langle v,\ y \rangle \Big\} \\ \text{subject to } p \geq 0, \text{and} & \, p_i \geq \frac{1}{2\rho} (\langle x_i,\ v \rangle)^2 - \lambda \quad \text{for } i = 1, \ldots, d. \end{aligned}$$

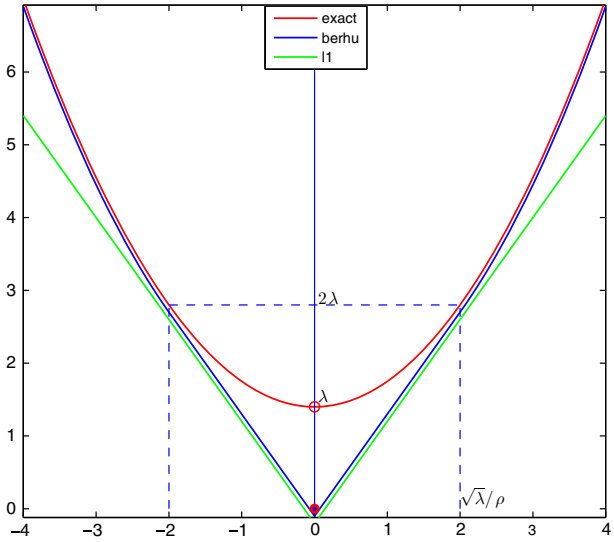

Fig. 2 Plots of three different penalty functions as a function of $t \in \mathbb{R}$: reverse Huber (berhu) function $t \mapsto B(\frac{\sqrt{\rho}t}{\lambda})$ ℓ_1-norm $t \mapsto \lambda|t|$ and the ℓ_0-based penalty $t \mapsto \frac{t^2}{2} + \lambda\|t\|_0$

Making use of rotated second order cone constraints, we have the equivalence

$$p_i \geq \frac{1}{2\rho}\big(\langle x_i,\ v\rangle\big)^2 - \lambda \iff \left\|\begin{pmatrix} \langle x_i,\ v\rangle \\ p_i + \lambda - 1 \end{pmatrix}\right\| \leq p_i + \lambda + 1, \qquad \text{for } i = 1, \ldots, d.$$

Thus, the relaxation (29) has the equivalent representation

$$\begin{aligned} P_{\text{IR}}(\lambda) = \max_{\substack{v\in\mathbb{R}^n \\ p\in\mathbb{R}^d}} & \left\{ -\langle 1,\ p\rangle - \frac{1}{2}\|v\|_2^2 - \langle v,\ y\rangle \right\} \\ \text{subject to} \quad & p \geq 0, \quad \left\|\begin{pmatrix} \sqrt{\rho^{-1}}\langle x_i,\ v\rangle \\ p_i + \lambda - 1 \end{pmatrix}\right\| \leq p_i + \lambda + 1, \quad i = 1, \ldots, d, \end{aligned}$$

which is a second order cone program (SOCP) in variables $(v, p) \in \mathbb{R}^n \times \mathbb{R}^d$.

Introducing Lagrange vectors for the constraints, we have

$$\begin{aligned} P_{\text{IR}}(\lambda) = \max_{v,p}\ \min_{\alpha,\beta,\gamma} & \left\{ -\langle 1,\ p\rangle - \frac{1}{2}\|v\|_2^2 - \langle v,\ y\rangle + \sum_{i=1}^{d} \big(\gamma_i(p_i + \lambda - 1) \right. \\ & \left. -\sqrt{\rho^{-1}}\alpha_i\langle x_i,\ v\rangle - \beta_i(p_i + \lambda + 1)\big) \right\} \\ \text{subject to} \quad & p \geq 0, \quad \left\|\begin{pmatrix} \alpha_i \\ \beta_i + \lambda - 1 \end{pmatrix}\right\| \leq \gamma_i, \quad i = 1, \ldots, d\,. \end{aligned}$$

Since $\lambda > 0$, strong duality holds by primal strict feasibility (see e.g., [4]), we may exchange the order of the minimum and the maximum. Making the substitutions $w = \alpha/\rho, u = \gamma + \beta, z = \gamma - \beta$, and then eliminating $v = \mathbf{y} - Xw$ yields the equivalent expression

$$\begin{aligned}
P_{\text{IR}}(\lambda) &= \min_{w,u,z} \max_{p \geq 0} \left\{ \frac{1}{2}\|Xw - y\|_2^2 + \langle p,\ z - 1\rangle + \langle 1,\ \lambda z + y\rangle \right\} \\
&\qquad \text{subject to} \quad \left\| \begin{pmatrix} \sqrt{\rho} x_i \\ y_i - z_i \end{pmatrix} \right\| \leq y_i + z_i \quad i = 1, \ldots, n. \\
&= \min_{w,u,z} \left\{ \frac{1}{2}\|Xw - y\|_2^2 + \langle p,\ \lambda z + y\rangle \right\} \\
&\qquad \text{subject to} \quad 0 \leq z_i \leq 1,\ y_i \geq 0,\ \rho w_i^2 \leq y_i z_i, \quad i = 1, \ldots, n \\
&= \min_{w,z} \left\{ \frac{1}{2}\|Xw - y\|_2^2 + \sum_{i=1}^{d} \left(\lambda z_i + \frac{\rho w_i^2}{z_i}\right) \right\}, \quad 0 \leq z_i \leq 1,\ i = 1, \ldots, n\,, \\
&= \min_{w} \left\{ \frac{1}{2}\|Xw - y\|_2^2 + 2\lambda \sum_{i=1}^{d} B\left(\frac{\sqrt{\rho} w_i}{\sqrt{\lambda}}\right) \right\},
\end{aligned}$$

which completes the proof. □

We note that the alternative reverse Huber representation of the least squares problem can potentially be used to apply convex optimization toolboxes (e.g., [8,15]) where the reverse Huber function is readily available.

5 Numerical results

In this section, we discuss some numerical aspects of solving the relaxations that we have introduced, and illustrate their behavior on some real-world problems of sparse learning.

5.1 Optimization techniques

Although efficient polynomial-time methods exist for solving semi-definite programs, solving large-scale problems remains challenging using current computers and algorithms. For the SDP problems of interest here, one attractive alternative is to instead develop algorithms to solve the saddle-point problem in Eq. (10). For instance, in the least-squares case, the gradients of the relaxed objective in Eq. (19) are given by

$$\partial_i G(u) = -\left(x_i^T (I + \mathrm{X} D(u) \mathrm{X}^T/\rho)^{-1} y\right)^2.$$

Computing such a gradient requires the solution of a rank-$\|u\|_0$ linear system of size n, which can be done exactly in time $\mathcal{O}(\|u\|_0^3) + \mathcal{O}(nd)$ via the QR decomposition. Therefore, the overall complexity of using first-order and quasi-Newton methods is

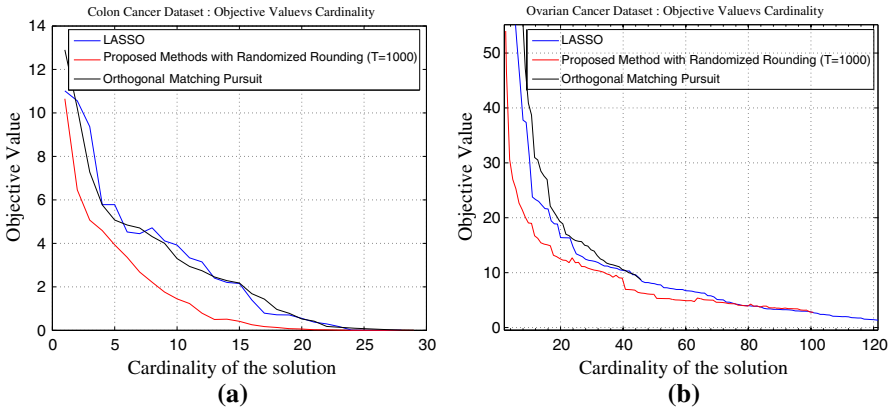

Fig. 3 Objective value versus cardinality trade-off in a real dataset from cancer research. The proposed randomized rounding method considerably outperforms other methods by achieving lower objective value with smaller cardinality. **a** Colon cancer dataset. **b** Ovarian cancer dataset

comparable to the Lasso when the sparsity level k is relatively small. We then employ a projected quasi-Newton method [28] to numerically optimize the convex objective. The randomized rounding procedure requires T evaluations of function value, which takes additional $\mathcal{O}(T\|\widetilde{u}\|_0^3)$ time.

5.2 Experiments on real datasets

We consider two well known high-dimensional datasets studied in cancer research, the 62×2000 *Colon cancer* dataset[1] and 216×4000 *Ovarian cancer* dataset[2] which contain ion intensity levels corresponding to related proteins and corresponding *cancer* or *normal* output labels. We consider classical ℓ_2^2-regularized least lquares classification using the mapping -1 for *cancer* label and $+1$ for *normal* label. We numerically implemented the proposed randomized rounding procedure of $T = 1000$ trials based on the relaxed solution. For other methods we identify their support and predict using regularized least squares solution constrained to that support where regularization parameter is optimized for each method on the training set. Figure 3 depicts optimization error (training error) as a function of the cardinality of the solution for both of the datasets. It is observed that the randomized rounding approach provides a considerable improvement in the optimal value for any fixed cardinality. In order to assess the learning and generalization performance of the trained model, we then split the dataset into two halves for training and testing. We present the plots of the test error as a function of cardinality over 1000 realizations of data splits and show the corresponding errorbars calculated for 1.5σ in Fig. 4. The proposed algorithm also shows a considerable

[1] Taken from the Princeton University Gene Expression Project; for original source and further details please see the references therein.

[2] Taken from FDA-NCI Clinical Proteomics Program Databank; for original source and further details please see the references therein.

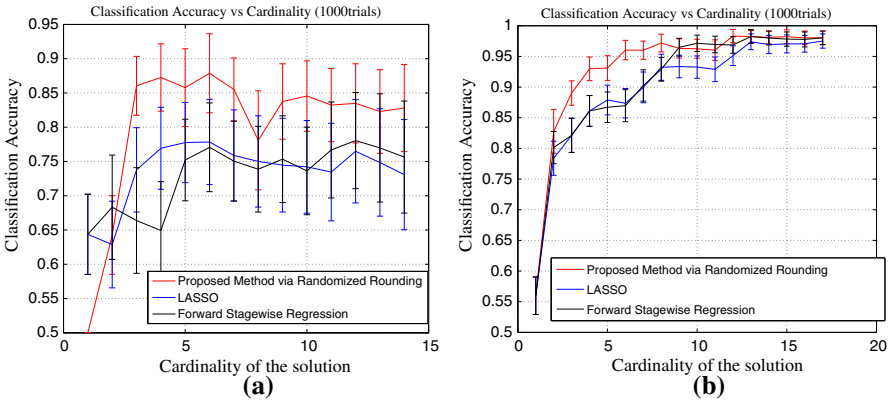

Fig. 4 Classification accuracy versus cardinality in a real dataset from cancer research. The proposed method has considerably higher classification accuracy for a fixed cardinality. **a** Colon cancer dataset. **b** Ovarian cancer dataset

improvement in both training and test error compared to the other methods, as can be seen from the figures. We observed that choosing $T \in [100, 1000]$ gave satisfactory results however T can be chosen larger for higher dimensional problems without any computational difficulty.

We also note that in many applications choosing a target cardinality k with good predictive accuracy is an important problem. For a range of cardinality values the proposed approach can be combined with cross-validation and other model selection methodologies such as the Bayesian information criterion (BIC) or Akaike information criterion (AIC) [2,36]. However there are also machine learning applications where the target cardinality is specified due to computational complexity requirements at runtime (see e.g. [11]). In these applications the cardinality directly effects the number of features that needs to be checked for classifying a new sample.

6 Discussion

In this paper, we first showed how a broad class of cardinality-constrained (or penalized) sparse learning problems can be reformulated exactly as Boolean programs involving convex objective functions. The utility of this reformulation is in permitting the application of various types of relaxation hierarchies, such as the Sherali-Adams and Lasserre hierarchies for Boolean programs. The simplest such relaxation is the first-order interval relaxation, and we analyzed the conditions for its exactness in detail. In contrast to the classical ℓ_1 heuristic, the presented method provides a lower bound on the solution value, and moreover a certificate of optimality when the solution is integral. We provided sufficient conditions for the solution to be integral for linear regression problems with random Gaussian design matrices. For problems in which the solution is not integral, we proposed an efficient randomized rounding procedure, and showed that its approximation accuracy can be controlled in terms of the number of fractional entries, and a regularization parameter in the algorithm, In our experiments

with real data sets, the output of this randomized rounding procedure provided considerably better solutions than standard competitors such as the Lasso or orthogonal matching pursuit.

There are a range of interesting open problem suggested by this paper. In particular, we have studied only the most naive first-order relaxation for the problem: it would be interesting to see whether one quantify how quickly the performance improves (relative to the exact cardinality-constrained solution) as the level of relaxation—say in one of the standard hierarchies for Boolean problems [16,17,19,29,35]—is increased. This question is particularly interesting in light of recent work [37] showing that, under a standard conjecture in computational complexity, there are fundamental gaps between the performance of cardinality-constrained estimators and polynomial-time methods for the prediction error in sparse regression.

Acknowledgments Authors MP and MJW were partially supported by Office of Naval Research MURI grant N00014-11-1-0688, and National Science Foundation Grants CIF-31712-23800 and DMS-1107000. In addition, MP was supported by a Microsoft Research Fellowship.

7 Appendix: Proofs

In this appendix, we provide the proofs of Theorems 2 and 3.

7.1 Proof of Theorem 2

Recalling the Definition (20) of the matrix M, for each $j \in \{1, \ldots, d\}$, define the rescaled random variable $U_j := \frac{X_j^T M y}{\rho n}$. In terms of this notation, it suffices to find a scalar λ such that

$$\min_{j \in S} |U_j| > \lambda \quad \text{and} \quad \max_{j \in S^c} |U_j| < \lambda. \tag{32}$$

By definition, we have $y = X_S w_S^* + \varepsilon$, whence

$$U_j = \underbrace{\frac{X_j^T M X_S w_S^*}{\rho n}}_{A_j} + \underbrace{\frac{X_j^T M \varepsilon}{\rho n}}_{B_j}.$$

Based on this decomposition, we then make the following claims:

Lemma 1 *There are numerical constants c_1, c_2 such that*

$$\mathbb{P}\Big[\max_{j=1,\ldots,d} |B_j| \geq t\Big] \leq c_1 e^{-c_2 \frac{nt^2}{\gamma^2} + \log d}. \tag{33}$$

Lemma 2 *There are numerical constants c_1, c_2 such that*

$$\mathbb{P}\Big[\min_{j\in S}|A_j| < \frac{w_{min}}{4}\Big] \leq c_1 e^{-c_2 n \frac{w_{min}^2}{\|w_S^*\|_2^2} + \log(2k)} \quad \textit{and} \tag{34a}$$

$$\mathbb{P}\Big[\max_{j\in S^c}|A_j| \geq \frac{w_{min}}{16}\Big] \leq c_3 e^{-c_4 n \frac{w_{min}^2}{\|w_S^*\|_2^2} + \log(d-k)}, \tag{34b}$$

Using these two lemmas, we can now complete the proof. Recall that Theorem 2 assumes a lower bound of the form $n > c_0 \frac{\gamma^2 + \|w_S^*\|_2^2}{w_{\min}^2} \log d$, where c_0 is a sufficiently large constant. Thus, setting $t = \frac{w_{\min}}{16}$ in Lemma 1 ensures that $\max_{j=1,\ldots,d}|B_j| \leq \frac{w_{\min}}{16}$ with high probability. Combined with the bound (34a) from Lemma 2, we are guaranteed that

$$\min_{j\in S}|U_j| \geq \frac{w_{\min}}{4} - \frac{w_{\min}}{16} = \frac{3w_{\min}}{16} \qquad \text{with high probability.}$$

Similarly, the bound (34b) guarantees that

$$\max_{j\in S^c}|U_j| \leq \frac{w_{\min}}{16} + \frac{w_{\min}}{16} = \frac{2w_{\min}}{16} \qquad \text{also with high probability.}$$

Thus, setting $\lambda = \frac{5w_{\min}}{32}$ ensures that the condition (32) holds.

The only remaining detail is to prove the two lemmas.

Proof of Lemma 1 Define the event $\mathcal{E}_j = \{\|X_j\|_2/\sqrt{n} \leq 2\}$, and observe that

$$\mathbb{P}\Big[|B_j| > t\Big] \leq \mathbb{P}[|B_j| > t \mid \mathcal{E}] + \mathbb{P}[\mathcal{E}^c].$$

Since the variable $\|X_j\|_2^2$ follows a χ^2-distribution with n degrees of freedom, we have $\mathbb{P}\big[\mathcal{E}^c\big] \leq 2e^{-c_2 n}$. Recalling the Definition (20) of the matrix M, note that $\sigma_{\max}\big(M\big) \leq \rho^{-1}$, whence conditioned on $\mathcal{E}$, we have $\|MX_j\|_2 \leq \|X_j\|_2 \leq 2\sqrt{n}$. Consequently, conditioned on $\mathcal{E}$, the variable $\frac{X_j^T M\varepsilon}{\rho}$ is a Gaussian random vector with variance at most $4\gamma^2/\rho^2$, and hence $\mathbb{P}[|B_j| > t \mid \mathcal{E}] \leq 2e^{-\frac{\rho^2 t^2}{32\gamma^2}}$.

Finally, by union bound, we have

$$\mathbb{P}\Big[\max_{j=1,\ldots,d}|B_j| > t\Big] \leq d\,\mathbb{P}\Big[|B_j| > t\Big] \leq d\Big\{2e^{-\frac{\rho^2 t^2}{32\gamma^2}} + 2e^{-c_2\rho n}\Big\} \leq c_1 e^{-c_2\frac{\rho^2 t^2}{\gamma^2} + \log d},$$

as claimed. □

Proof of Lemma 2 We split the proof into two parts.

(1) Proof of the bound (34a):
Note that

$$\frac{1}{\rho} X_S^T M X_S = X_S^T (\rho I_n + X_S X_S^T)^{-1} X_S$$

We now write $X_S = U D V^T$ for singular value decomposition of $\frac{1}{\sqrt{n}} X_S$ in compact form. We thus have

$$\frac{1}{\rho} X_S^T M X_S = V \left(\rho I_n + n D^2\right)^{-1} D^2 V^T.$$

We will prove that for a fixed vector z, the following holds with high probability

$$\frac{\| \left(\frac{1}{\rho} X_S^T M X_S - I\right) z \|_\infty}{\|z\|_\infty} \leq \epsilon. \tag{35}$$

Applying the above bound to w_S^*, which is a fixed vector we obtain

$$\| \left(\frac{1}{\rho} X_S^T M X_S - I\right) w_s^* \|_\infty \leq \epsilon \|w_s^*\|_\infty \tag{36}$$

Then by triangle inequality the above statement implies that

$$\min_{i \in S} |\frac{1}{\rho} X_S^T M X_S w_i^*| > (1 - \epsilon) \min_{i \in S} |w_i^*|.$$

and setting $\epsilon = 3/4$ yields the claim.

Next we let $\frac{1}{\rho} X_S^T M X_S - I = V \tilde{D} V$ where we defined $\tilde{D} := \left((\rho I_n + D^2)^{-1} D^2 - I\right)$. By standard results on operator norm of Gaussian random matrices (e.g., see Davidson and Szarek [10]), the minimum singular valyue

$$\sigma_{\min} \left(\frac{1}{\sqrt{n}} \mathrm{X}_S\right) = \min_{i=1,\ldots,k} D_{ii}$$

of the matrix $\mathrm{X}_S/\sqrt{n}$ can be bounded as

$$\mathbb{P}\left[\frac{1}{\sqrt{n}} \min_{i=1,\ldots,k} |D_{ii}| \leq 1 - \sqrt{\frac{k}{n}} - t\right] \leq 2e^{-c_1 n t^2}, \tag{37}$$

where c_1 is a numerical constant (independent of (n, k)).

Now define $Y_i := e_i^T V \tilde{D} V^T z = z_i v_i \tilde{D} v_i + v_i^T \tilde{D} \sum_{l \neq i} z_l v_l$. Then note that,

$$|Y_1| \leq \|\tilde{D}\|_2 |z_1| + v_1^T \tilde{D} \sum_{l \neq i} z_l v_l = \frac{\rho}{\rho + \min_{i=1,\ldots,k} |D_{ii}|^2} |z_1| + F(v_1)$$

where we defined $F(v_1) := v_1^T \tilde{D} \sum_{l \neq i} z_l v_l$ and v_1 is uniformly distributed over a sphere in $k-1$ dimensions and hence $\mathbb{E} F(v_1) = 0$. Observe that F is a Lipschitz map satisfying

$$\begin{aligned} |F(v_1) - F(v_1')| &\leq \|\tilde{D}\|_\infty \sqrt{\sum_{l \neq i} |z_l^2|} |v_1 - v_1'\|_2 \\ &= \frac{\rho}{\rho + \min_i |D_{ii}|^2} |\sqrt{k-1}\|z\|_\infty \|v_1 - v_1'\|_2 \end{aligned}$$

Applying concentration of measure for Lipschitz functions on the sphere (e.g., see [18]) the function $F(v_1)$ we get that for all $t > 0$ we have,

$$\mathbb{P}\big[F(v_1) > t\|z\|_\infty\big] \leq 2e^{-c_4(k-1)\frac{t^2}{\left(\frac{\rho}{\rho + \min_i |D_{ii}|^2}\right)^2 (k-1)}}. \tag{38}$$

Conditioning on the high probability event $\{\min_i |D_{ii}|^2 \leq \frac{n}{2}\}$ and then applying the tail bound (37) yields

$$\begin{aligned} \mathbb{P}\big[F(v_1) > t\|z\|_\infty\big] &\leq 2\exp\left(-c_4 \frac{n^2 t^2}{\rho^2}\right) + 2e^{-c_2 \frac{nt^2}{\rho^2}} \\ &\leq 4e^{-c_5 \frac{n^2 t^2}{\rho^2}}. \end{aligned} \tag{39}$$

Combining the pieces in (39) and (38), we take a union bound over $2k$ coordinates,

$$\begin{aligned} \mathbb{P}\left[\min_{j \in S} |Y_j| > t\|z\|_\infty\right] &\leq 2k\ 3\exp\left(-c_5 n^2 t^2/\rho^2\right) \\ &\leq 2k\ 3\exp\left(-c_5 n t^2\right). \end{aligned}$$

where the final line follows from our choice $\rho = \sqrt{n}$. Finally setting $t = \epsilon$ we obtain the statement in (35) and hence complete the proof.

Proof of the bound (34b): A similar calculation yields

$$A_j = \frac{1}{\rho} X_{S^c}^T M X_S w_S^* = X_{S^c}^T \big(\rho I_n + X_S X_S^T\big)^{-1} X_s w_S^*,$$

for each $j \in S^c$.

Defining the event $\mathcal{E} = \{\sigma_{\max}(X_S)/ \leq 2\sqrt{n}\}$, standard bounds in random matrix theory [10] imply that $\mathbb{P}[\mathcal{E}^c] \leq 2e^{-c_2 n}$. Conditioned on $\mathcal{E}$, we have

$$\|\big(\rho I_n + X_S X_S^T\big)^{-1} X_s w_S^*\|_2 \leq \frac{2}{\rho}\|w_S^*\|_2,$$

so that the variable A_j is conditionally Gaussian with variance at most $\frac{4}{\rho^2}\|w_S^*\|_2^2$. Consequently, we have

$$\mathbb{P}[|A_j| \geq t] \leq \mathbb{P}[|A_j| \geq t \mid \mathcal{E}] + \mathbb{P}[\mathcal{E}^c] \;=\; 2e^{-\frac{\rho^2 t^2}{32\|w_S^*\|_2^2}} + 2e^{-c_2} \leq c_1 e^{-c_2 \frac{\rho^2 t^2}{\|w_S^*\|_2^2}},$$

Setting $t = \frac{w_{\min}}{8}$, $\rho = \sqrt{n}$ and taking union bound over all $d - k$ indices in S^c yields the claim (34b). □

7.2 Proof of Theorem 3

The vector $\widetilde{u} \in \{0,1\}^d$ consists of independent Bernoulli trials, and we have $\mathbb{E}[\sum_{j=1}^d \widetilde{u}_j] \leq k$. Consequently, by the Chernoff bound for Bernoulli sums, we have

$$\mathbb{P}\Big[\sum_{j=1}^d \widetilde{u}_j \geq (1+\delta)k\Big] \leq c_1 e^{-c_2 k\delta^2}.$$

as claimed.

It remains to establish the high-probability bound on the optimal value. As shown previously, the Boolean problem admits the saddle point representation

$$P^* = \min_{u\in\{0,1\}^d,\ \sum_{i=1}^d u_i \leq k} \Big\{ \underbrace{\max_{\alpha\in\mathbb{R}^n} -\frac{1}{\rho}\alpha^T \mathrm{X} D(u) \mathrm{X}^T \alpha - \|\alpha\|_2^2 - 2\alpha^T y}_{G(u)} \Big\}. \tag{40}$$

Since the optimal value is non-negative, the optimal dual parameter $\alpha \in \mathbb{R}^n$ must have its ℓ_2-norm bounded as $\|\alpha\|_2 \leq 2\|y\|_2 \leq 2$. Using this fact, we have

$$\begin{aligned} G(\widehat{u}) - G(\widetilde{u}) &= \max_{\|\alpha\|_2\leq 2} \Big\{ -\frac{1}{\rho}\alpha^T \mathrm{X} D(\widehat{u}) \mathrm{X}^T \alpha - \|\alpha\|_2^2 - 2\alpha^T y \Big\} \\ &\qquad - \max_{\|\alpha\|_2\leq 2} \Big\{ -\frac{1}{\rho}\alpha^T \mathrm{X} D(\widetilde{u}) \mathrm{X}^T \alpha - \|\alpha\|_2^2 - 2\alpha^T y \Big\} \\ &\leq \max_{\|\alpha\|_2\leq 2} \Big\{ -\frac{1}{\rho}\alpha^T \mathrm{X}(D(\widehat{u}) - D(\widetilde{u})) \mathrm{X}^T \alpha \Big\} \\ &\leq \frac{2}{\rho}\sigma_{\max}\big(\mathrm{X}(D(\widehat{u}) - D(\widetilde{u}))\mathrm{X}^T\big), \end{aligned}$$

where $\sigma_{\max}(\,\cdot\,)$ denotes the maximum eigenvalue of a symmetric matrix.

It remains to establish a high probability bound on this maximum eigenvalue. Recall that R is the subset of indices associated with fractional elements of $\widehat{u}$, and moreover that $\mathbb{E}[\widetilde{u}_j] = \widehat{u}_j$. Using these facts, we can write

$$\mathrm{X}(D(\widetilde{u}) - D(\widehat{u}))\mathrm{X}^T = \sum_{j \in R} \underbrace{\big(\widetilde{u}_j - \mathbb{E}[\widetilde{u}_j]\big) X_j X_j^T}_{A_j}$$

where $X_j \in \mathbb{R}^n$ denotes the jth column of X. Since $\|X_j\|_2 \leq 1$ by assumption and $\widetilde{u}_j$ is Bernoulli, the matrix A_j has operator norm at most 1, and is zero mean. Consequently, by the Ahlswede-Winter matrix bound [1,26], we have

$$\mathbb{P}\Big[\sigma_{\max}\big(\sum_{j \in R} A_j\big) \geq \sqrt{r}t\Big] \leq 2\min\{n, r\}e^{-t^2/16},$$

where $r = |R|$ is the number of fractional components. Setting $t^2 = c \log \min\{n, r\}$ for a sufficiently large constant c yields the claim.

References

1. Ahlswede, R., Winter, A.: Strong converse for identification via quantum channels. IEEE Trans. Inf. Theory **48**(3), 569–579 (2002)
2. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Proceedings of the 2nd international symposium on information theory, Tsahkadsor, Armenia, USSR (September 1971)
3. Bickel, P.J., Ritov, Y., Tsybakov, A.: Simultaneous analysis of Lasso and Dantzig selector. Ann. Stat. **37**(4), 1705–1732 (2009)
4. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge, UK (2004)
5. Bühlmann, P., van de Geer, S.: Statistics for High-Dimensional Data. Springer Series in Statistics. Springer, Berlin (2011)
6. Candes, E.J., Tao, T.: Decoding by linear programming. IEEE Trans. Info. Theory **51**(12), 4203–4215 (2005)
7. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines (and Other Kernel Based Learning Methods). Cambridge University Press, Cambridge (2000)
8. Inc. CVX Research. CVX: Matlab software for disciplined convex programming, version 2.0 (August 2012)
9. d'Aspremont, A., El Ghaoui, L.: Testing the nullspace property using semidefinite programming. Technical report, Princeton (2009)
10. Davidson, K.R., Szarek, S.J.: Local operator theory, random matrices and Banach spaces. Handbook of Banach Spaces, vol. 1, pp. 317–336. Elsevier, Amsterdam (2001)
11. Dekel, O., Singer, Y.: Support vector machines on a budget. Adv. Neural Inf. Process. Syst. **19**, 345 (2007)
12. Donoho, D.L.: Compress. sensing. IEEE Trans. Info. Theory **52**(4), 1289–1306 (2006)
13. Donoho, D.L., Elad, M., Temlyakov, V.M.: Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Trans. Inf. Theory **52**(1), 6–18 (2006)
14. Fuchs, J.J.: Recovery of exact sparse representations in the presence of noise. ICASSP **2**, 533–536 (2004)
15. Grant, M., Boyd, S.: Graph implementations for nonsmooth convex programs. In: Blondel, V., Boyd, S., Kimura, H. (eds.) Recent Advances in Learning and Control, Lecture Notes in Control and Information Sciences, pp. 95–110. Springer, Berlin (2008)

16. Lasserre, J.B.: An explicit exact SDP relaxation for nonlinear 0–1 programs. In: Aardal K., and Gerads A.M.H., (eds.) Lecture Notes in Computer Science, 2081:293–303 (2001)
17. Laurent, M.: A comparison of the Sherali-Adams, Lovász-Schrijver and Lasserre relaxations for 0–1 programming. Math. Oper. Res. **28**, 470–496 (2003)
18. Ledoux, M.: The Concentration of Measure Phenomenon. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI (2001)
19. Lovász, L., Schrijver, A.: Cones of matrices and set-functions and 0–1 optimization. SIAM J. Optim. **1**, 166–190 (1991)
20. Markowitz, H.M.: Portf. Sel. Wiley, New York (1959)
21. McCullagh, P., Nelder, J.A.: Generalized Linear Models. Monographs on Statistics and Applied Probability 37. Chapman and Hall/CRC, New York (1989)
22. Meinshausen, N., Bühlmann, P.: High-dimensional graphs and variable selection with the Lasso. Ann. Stat. **34**, 1436–1462 (2006)
23. Motwani, R., Raghavan, P.: Randomized Algorithms. Cambridge University Press, Cambridge, UK (1995)
24. Negahban, S., Ravikumar, P., Wainwright, M.J., Yu, B.: Restricted strong convexity and generalized linear models. Technical report, UC Berkeley, Department of Statistics (August 2011)
25. Nesterov, Y.: Primal-dual subgradient methods for convex problems. Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL) (2005)
26. Oliveira, R.I.: Sums of random Hermitian matrices and an inequality by Rudelson. Elec. Comm. Prob. **15**, 203–212 (2010)
27. Pilanci, M., El Ghaoui, L., Chandrasekaran, V.: Recovery of sparse probability measures via convex programming. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 2420–2428. Curran Associates, Inc. (2012)
28. Schmidt, M., van den Berg, E., Friedlander, M., Murphy, K.: Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm. AISTATS 2009, 5 (2009)
29. Sherali, H.D., Adams, W.P.: A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. SIAM J. Discrete Math. **3**, 411–430 (1990)
30. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B **58**(1), 267–288 (1996)
31. Tropp, J.A.: Just relax: Convex programming methods for subset selection and sparse approximation. ICES Report 04–04, UT-Austin, February (2004)
32. Wainwright, M.J.: Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting. IEEE Trans. Info. Theory **55**, 5728–5741 (2009)
33. Wainwright, M.J.: Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). IEEE Trans. Inf. Theory **55**, 2183–2202 (2009)
34. Wainwright, M.J.: Structured regularizers: statistical and computational issues. Annu. Rev. Stat. Appl. **1**, 233–253 (2014)
35. Wainwright, M.J., Jordan, M.I.: Treewidth-based conditions for exactness of the Sherali-Adams and Lasserre relaxations. Technical report, UC Berkeley, Department of Statistics, No. 671 (September 2004)
36. Wasserman, Larry: Bayesian model selection and model averaging. J. Math. Psychol. **44**(1), 92–107 (2000)
37. Zhang, Y., Wainwright, M.J., Jordan, M.I.: Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In COLT conference, Barcelona, Spain, (June 2014). Full length version at http://arxiv.org/abs/1402.1918
38. Zhao, P., Yu, B.: On model selection consistency of Lasso. J. Mach. Learn. Res. **7**, 2541–2567 (2006)
39. Zou, H., Hastie, T.J.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B **67**(2), 301–320 (2005)