Noname manuscript No.
(will be inserted by the editor)

Regularization vs. Relaxation: A convexification perspective of statistical variable selection

Hongbo Dong · Kun Chen · Jeff Linderoth

Initial version: 05/28/2015, current version: March 4, 2018

Abstract Variable selection is a fundamental task in statistical data analysis. One central problem is to minimize a loss function with additional binary indicator functions modeling “sparsity", either in the objective or constraints. Different sparsity-inducing penalty functions are used to approximate these binary indicator functions in the literature. In this paper, we show that viewing the problem from a convexification perspective offers new insights, and potentially leads to a unified framework for deriving penalty functions. Under mild conditions, a convex reformulation (in exponential size) of the original nonconvex problem is obtained by disjunctive techniques. A special case of the disjunctive convexification is known as perspective relaxation in mixed-integer nonlinear optimization. We derive their equivalent penalty forms (named as *perspective penalty functions*). We show that a popular sparsity-inducing concave penalty function known as the Minimax Concave Penalty (MCP), and the reverse Huber penalty derived in a recent work by Pilanci, Wainwright and El Ghaoui, can both be seen as special cases of the perspective penalty functions. Finally we study the “optimal perspective relaxation" with a minimax formulation. This relaxation provides valid bounds of optimal values in the original nonconvex problem. We show the existence and attainment of saddle points for the case of general (strongly) convex loss functions. When the loss function is (strongly) convex quadratic, we show it can be solved by a semidefinite program. An interesting fact is that the proposed semidefinite relaxation can be realized as a well-known semidefinite relaxation for boolean quadratic programming embedded into a two-level formulation of

The second author is partially supported by National Science Foundation grants DMS-1613295 and IIS-1718798; the third author is supported in part by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under contract number DE-AC02-06CH11357.

Hongbo Dong
Department of Mathematics, Washington State University, Pullman, WA 99163
E-mail: hongbo.dong@wsu.edu

Kun Chen
Department of Statistics, University of Connecticut, Storrs, CT 06269
E-mail: kun.chen@uconn.edu

Jeff Linderoth
Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI 53706
E-mail: linderoth@wisc.edu

the original problem with binary indicator functions. This suggests using Goemans-Williamson rounding procedure to find approximate solutions. Computational results are provided in both low and high dimensional settings to illustrate improvements in bound tightness over previous convex relaxations.

Keywords Sparse optimization · convex relaxation · semidefinite programming · minimax concave penalty

Mathematics Subject Classification 90C22, 90C47, 62J07

1 Introduction

We consider the following optimization problem with binary indicator functions, that is fundamental in variable selection in statistical learning, and compressive sensing,

$$\zeta_{L0} := \min_{\beta} \ \mathscr{L}(\beta) + \lambda \sum_{j \in \mathscr{I}} |\beta_j|_0, \quad s.t. \quad \sum_{j \in \mathscr{I}} |\beta_j|_0 \le K. \tag{L_0}$$

where $\mathscr{L}(\bullet) : \mathbb{R}^p \mapsto \mathbb{R}$ is a convex loss function, and $|\beta_j|_0$ is the binary indicator function of β_j in the sense that $|\beta_j|_0 = 0$ if $\beta_j = 0$ and $|\beta_j|_0 = 1$ otherwise. $\mathscr{I}$ is a subset of $\{1, ..., p\}$. $\lambda \ge 0$ and K are both parameters controlling the number of nonzero entries in solution.

The penalized optimization approaches, capable of simultaneous dimension reduction and model estimation, have undergone exciting developments in the statistical community in recent years. These approaches typically use continuous functions $\rho(\beta_j)$ to approximate $|\beta_j|_0$. The design of penalty functions, optimization algorithms for solving such approximations, and the properties of the resulting estimators have been extensively studied in the statistical literature. Popular methods include the lasso [37], the adaptive lasso [43,29], the group lasso [40], the elastic net [44,46], the smoothly clipped absolute deviation (SCAD) penalty [15], the bridge regression [22,28], the minimax concave penalty (MCP) [41] and the smooth integration of counting and absolute deviation (SICA) penalty [32,17]. Algorithms have been developed to solve the lasso problem and its variants, e.g. the least angle regression algorithm [14] and the coordinate descent algorithm [38,23]. For optimizing a nonconvex penalized likelihood, Fan and Li proposed an iterative local quadratic approximation (LQA). Zou and Li in [45] developed an iterative algorithm based on local linear approximation (LLA), which was shown to be the best convex minorization-maximization (MM) algorithm [30]. These local approximation approaches are commonly coupled with coordinate descent to solve general penalized likelihood problems [39,9]. For a comprehensive account of these approaches from a statistical perspective, see [10], [16] and [27]. Note that such penalty functions are usually designed to approximate the binary indicator function $|\cdot|_0$, while rarely exploit structures in $\mathscr{L}(\bullet)$.

Exact and approximate methods for solving (L_0) also received much attention from the optimization community in recent years. One approach is to directly solve (L_0) by using tailored branch-and-bound algorithms pioneered in [8,6]. Recently, Bertsimas, King, and Mazumder [5] showed that with properly-engineered techniques from mixed-integer quadratic programming, (L_0) with convex quadratic $\mathscr{L}(\bullet)$ can be solved approximately or exactly for some instances of practical size. The authors in [19] show promising computational results by formulating (L_0) as a nonlinear program with complementarity conditions, and use nonlinear optimization algorithms to find good feasible solutions. Pilanci, Wainwright and El Ghaoui [34] reformulated (L_0) into a convex nonlinear optimization problem with binary variables. By relaxing the binary conditions they obtained tractable convex relaxations for (L_0) (in the sense that optimal value of the relaxation is a lower bound of ζ_{L0}). Tractable lower bounds are useful in quantifying the quality of feasible solutions to (L_0).

The approach of our paper can be seen as an attempt of connecting related research in the optimization and statistical communities. Novel contributions of this paper are three-fold. Firstly, We provide a "complete" convexification of (L_0) based on its disjunctive formulation; such convexification is in general of exponential size, however can be tractable if $\mathscr{L}(\bullet)$ has a block-diagonal structure. Secondly, by decomposing $\mathscr{L}(\bullet)$ as the sum of a separable convex component and a non-separable component, and exploiting some low-dimensional convex hulls, we derive a class of penalty functions called "perspective penalties". This approach is a penalty counterpart of the "perspective relaxation" well known in mixed-integer nonlinear optimization [20,25,26]. Furthermore, the convex relaxation in [34] and the *minimax concave penalty* (MCP), a popular penalty function developed in the statistical community, can both be seen as special cases of perspective penalties. Thirdly, when $\mathscr{L}(\bullet)$ is strongly convex, perspective relaxations lead to non-trivial lower bound of ζ_{L0}. We focus on the "optimal perspective relaxation" in a minimax formulation. We show the existence and attainment of saddle points, and that when $\mathscr{L}(\bullet)$ is strongly convex quadratic, such a saddle point can be computed by semidefinite programming. An interesting fact is that the proposed semidefinite relaxation can be realized as a well-known semidefinite relaxation for boolean quadratic programming embedded into a two-level formulation of (L_0). This suggests using the Goemans-Williamson rounding procedure to find approximate solutions to (L_0). Computational results are presented to compare the semidefinite relaxation, the relaxation in [34] and branch-and-bound solvers with a time limit.

We explain some concepts in convex analysis and notation used in this paper. To avoid unnecessary complications we focus on functions taking values in $\mathbb{R}$ (as opposed to extended real values $\mathbb{R} \cup \{\pm\infty\}$) unless otherwise stated. For any function $f : \mathbb{R}^p \mapsto \mathbb{R}$, its *epigraph* is the set $\mathbf{epi}(f) := \{(x,t) \mid f(x) \le t\}$. For a non-empty convex set $C \subseteq \mathbb{R}^p$, its recession cone is $\mathscr{R}(C) := \{d \in \mathbb{R}^p \mid x + \lambda d \in C, \text{ for all } x \in C, \ \lambda \ge 0\}$. For a convex function $f : \mathbb{R}^p \mapsto \mathbb{R}$, its recession function $f0^+ : \mathbb{R}^p \mapsto \mathbb{R} \cup \{+\infty\}$ is a positive homogeneous proper convex function whose epigraph is the recession cone of $\mathbf{epi}(f)$, i.e.,

$$\mathbf{epi}(f0^+) = \mathscr{R}(\mathbf{epi}(f)).$$

The recession cone of f, denoted as $\mathscr{R}(f)$, which should not be confused with the recession cone of $\mathbf{epi}(f)$, is the set of all d such that $f0^+(d) \le 0$. Equivalently, $\mathscr{R}(f)$ is the recession cone of any *non-empty* level set of f. Any nonzero vector in $\mathscr{R}(f)$ is called a *recession direction* of f. For a convex function f : $\mathbb{R}^p \mapsto \mathbb{R}$, its perspective function is defined as

$$\tilde{f} : [0,+\infty) \times \mathbb{R}^n, \quad s.t., f(z,x) = \begin{cases} zf(x/z), & \text{for } z > 0, \\ f0^+(x), & \text{for } z = 0. \end{cases} \tag{1}$$

For any set $C \subseteq \mathbb{R}^p$, its convex hull is denoted as $\mathbf{conv}(C)$ while $\mathbf{clconv}(C)$ being its closure. The space of $p \times p$ real symmetric matrices is denoted by $\mathscr{S}^p$, and the space of $n \times p$ real matrices is denoted as $\mathbb{R}^{n\times p}$. The inner product between two matrices $A, B \in \mathbb{R}^{n\times p}$ is $\langle A,B\rangle = \mathbf{trace}(AB^T)$. Given a matrix $B \in \mathscr{S}^p$, we say $B \succ (\succeq) 0$ if it is positive (semi)definite. The cones of positive semidefinite matrices and positive definite matrices are denoted as $\mathscr{S}^p_+$ and $\mathscr{S}^p_{++}$, respectively. The matrix I is the identity matrix, and e is used to denote a vector with all entries equal 1, of a conformal dimension. For a vector $\delta \in \mathbb{R}^p$, $\mathbf{D}(\delta)$ is a $p \times p$ diagonal matrix whose diagonal entries are entries in δ. For a symmetric matrix B, $\mathbf{D}(B)$ is a vector of its diagonal entries. For a finite set $\mathscr{I}$, $|\mathscr{I}|$ is used to denote its cardinality.

We sometimes consider the following important class of loss functions: given a collection of samples $\{(x_i, y_i)\}_{i=1}^n$, where each $x_i \in \mathbb{R}^p$ is an observation of predictor variables and $y_i \in \mathbb{R}$ the corresponding outcome,

$$\mathscr{L}(\beta) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i^T\beta; y_i) + \frac{\mu}{2}\|\beta\|^2. \tag{2}$$

In the case of linear regression $\ell(t;\theta) = (t-\theta)^2/2$ while in logistical regression (where each $y_i \in \{0,1\}$) $\ell(t;\theta) = \log(1+exp(t)) - t\theta$. We use $X \in \mathbb{R}^{n\times p}$ to denote the matrix with i-th row x_i^T.

2 A Convexification approach

In this section we present a general approach for convexifying (L_0), where $\lambda \geq 0$ and K are nonnegative parameters controlling the solution sparsity. Convexification by disjunctive formulations has been studied for mixed-integer nonlinear optimization [36, 12]. However, due to the unboundedness of variables (in both positive and negative directions), results in these works do not immediately apply. We provide a fairly short derivation of disjunctive convex formulation for (L_0) in this section. We allow either $\lambda = 0$ or $K \geq |\mathscr{I}|$ while making the following assumption:

Assumption 1 $\mathscr{L}(\bullet) : \mathbb{R}^p \mapsto \mathbb{R}$ *is convex, bounded below, and at least one of the following two conditions holds:*

1. $\mathscr{R}(\mathscr{L}) = \{0\}$ *(or equivalently,* $\mathscr{L}(\bullet)$ *is level-bounded),*
2. $K < \min_d \left\{\sum_{j\in\mathscr{I}} |d_j|_0 \;\middle|\; d \in \mathscr{R}(\mathscr{L}),\ d \neq 0\right\}$.

Note that condition 2, which states that $\mathscr{L}$ has no sparse recession direction (except 0), is more general than condition 1. An implication is the compactness of global optimal solutions to (L_0):

Proposition 1 *Under assumption 1, the set of global optimal solutions to* (L_0) *is nonempty and compact.*

Proof Suppose $\mathscr{R}(\mathscr{L}) = \{0\}$, then the proof is immediate by the Weierstrass theorem for lower semicontinuous functions.

Suppose condition 2 in Assumption 1 holds, we claim that for any subset $J \subseteq \mathscr{I}$ and $|J| \leq K$, the restriction of $\mathscr{L}(\bullet)$ on the subspace $\{\beta \mid \beta_j = 0,\ \forall j \notin J\}$ is level-bounded. Let us denote $\mathscr{L}_J : \mathbb{R}^{|J|} \mapsto \mathbb{R}$ to be such a restriction. Then for all α, the level set $\left\{\beta_J \in \mathbb{R}^{|J|} \;\middle|\; \mathscr{L}_J(\beta_J) \leq \alpha\right\}$ has no direction of recession, as otherwise an embedding of such a direction to $\mathbb{R}^p$ implies a contradiction to condition 2 in Assumption 1 (this is again due to the convexity of $\mathscr{L}(\bullet)$, e.g., [35, Corollary 8.4.1]). Therefore $\mathscr{L}_J$ is level-bounded (e.g., see [35, Theorem 8.4]). Since any global optimal solution to (L_0) must correspond to a global solution to the restricted problem

$$\min_{\beta_J \in \mathbb{R}^{|J|}} \ \mathscr{L}_J(\beta_J) + \lambda \sum_{j\in J} |\beta_j|_0,$$

for some J, by the Weierstrass theorem for lower semicontinuous functions the global solution set to (L_0) is nonempty and bounded. Since the global solution set is also closed, it is compact. □

Remark 1 We show that Assumption 1 is reasonable and relates to popular assumptions in statistical learning. Consider the loss function $\mathscr{L}(\beta)$ in the form of (2), If $\mu > 0$, then obviously $\mathscr{R}(\mathscr{L}) = \{0\}$. Now consider the case of $\mu = 0$. In linear regression $\ell(t;\theta) = (t-\theta)^2/2$. By some calculus rules for recession functions (e.g., [35, Theorem 9.3 and 9.5]), one can compute that $\mathscr{R}(\mathscr{L})$ is the null space of X. Therefore condition 2 in Assumption 1 states that there exists no nonzero sparse vectors in the null space, a condition *weaker* than the sparse Riesz conditions (or restricted eigenvalue conditions) commonly used in the statistical learning literature e.g., [7] [41]. For the case of logistical regression (where each $y_i \in \{0,1\}$),

$$\ell 0^{+}(\bullet, y) = [t]_{+} - ty.$$

The recession cone of (2) with $\mu = 0$ is

$$\mathscr{R}(\mathscr{L}) = \left\{ d \in \mathbb{R}^p \;\middle|\; x_i^T d \geq 0,\ \forall y_i = 1,\ \ x_i^T d \leq 0,\ \forall y_i = 0 \right\}.$$

Therefore condition 2 in Assumption 1 implies the entire data set cannot be completely separated by a linear function of the predictor variables (with sparse coefficients). Otherwise it causes the “complete separation" issue known in logistical regression (e.g., [2]).

Now we present the general convexification approach by using disjunctive techniques. By introducing a binary vector $z \in \{0,1\}^{|\mathscr{I}|}$ to denote the sparsity pattern of $\{\beta_j\}_{j\in\mathscr{I}}$ [1], and a scalar variable $t \in \mathbb{R}$ to represent an upper bound of $\mathscr{L}(\beta)$, we can equivalently write (L_0) as

$$\zeta_{L0} = \min_{(t,\beta,z)} \ t + \lambda \sum_{j\in\mathscr{I}} z_j, \quad (t,\beta,z) \in \mathscr{F}(\mathscr{L},K), \qquad \text{where}$$

$$\mathscr{F}(\mathscr{L},K) := \left\{ (t,\beta,z) \in \mathbb{R}^{1+p} \times \{0,1\}^{|\mathscr{I}|} \;\middle|\; \mathscr{L}(\beta) \leq t,\ \beta_j(1-z_j) = 0,\ \sum_{j\in\mathscr{I}} z_j \leq K, \forall j \in \mathscr{I} \right\}. \tag{3}$$

By enumerating all sets $J \subseteq \mathscr{I}$ such that $|J| \leq K$, $\mathscr{F}(\mathscr{L},K)$ can be written as the following union:

$$\mathscr{F}(\mathscr{L},K) = \bigcup_{\substack{J\subseteq\mathscr{I}\\ |J|\leq K}} C^J, \quad \text{where} \ \ C^J := \left\{ (t,\beta,\mathbf{1}^J) \;\middle|\; \mathscr{L}(\beta) \leq t, \quad \beta^J_j = 0, \forall j \in \mathscr{I}\setminus J \right\}, \tag{4}$$

and $\mathbf{1}^J$ is a binary vector with entry 1 at all positions in J and 0 otherwise. A convexification for (L_0) is:

$$\zeta_{L0} = h^* := \min_{t,\beta,z} \quad t + \lambda \sum_{j\in\mathscr{I}} z_j, \quad s.t.,\ (t,\beta,z) \in \textbf{clconv}\mathscr{F}(\mathscr{L},K). \tag{L0-H}$$

We will use the following result in convex analysis to obtain an algebraic description for $\textbf{clconv}\mathscr{F}(\mathscr{L},K)$.

Theorem 1 (Theorem 9.8 in [35]) *Let $C_1,\ldots,C_m$ be non-empty closed convex sets in $\mathbb{R}^n$ satisfying the following condition: if $z_1,\ldots,z_m$ are vectors such that $z_i \in \mathscr{R}(C_i)$ and $z_1 + \cdots z_m = 0$, then each z_i is in the lineality space of C_i. We have*

$$\textbf{clconv}(C_1 \cup \cdots \cup C_m) = \left\{ \sum_{i=1}^m \mu_i x_i + \sum_{i=1}^m d_i \;\middle|\; \sum_{i=1}^m \mu_i = 1,\ \mu_i \geq 0, x_i \in C_i, d_i \in \mathscr{R}(C_i),\ \forall i = 1,\ldots,m \right\}.$$

Furthermore, it can be assumed that $d_i \neq 0$ only if $\mu_i = 0$.

[1] Naturally we index z such that z_j is the binary indicator for β_j, for all $j \in \mathscr{I}$.

Theorem 2 *Suppose that Assumption 1 holds,*

$$\mathbf{clconv}\mathscr{F}(\mathscr{L},K) := \left\{ (\sum_J t^J, \sum_J \beta^J, \sum_J \mu^J \mathbf{1}^J) \;\middle|\; \begin{array}{r} \tilde{\mathscr{L}}(\mu^J,\beta^J) \le t^J, \beta^J_j = 0, \quad \forall j \in \mathscr{I} \setminus J \\ \sum_J \mu^J = 1, \mu^J \ge 0, \quad \forall J \subseteq \mathscr{I}, |J| \le K \end{array} \right\},$$

where $\tilde{L}(\bullet,\bullet)$ is the perspective function of $\mathscr{L}(\bullet)$. If $\lambda > 0$, the set of optimal solutions to (L0-H) is

$$\mathbf{conv}\left\{ \left(\mathscr{L}(\bar{\beta}), \bar{\beta}, z\right) \;\middle|\; \bar{\beta} \textit{ is globally optimal to } (L_0),\ z \in \{0,1\}^{|\mathscr{I}|},\ z_j = |\beta_j|_0, \forall j \in \mathscr{I} \right\}. \tag{5}$$

Otherwise if $\lambda = 0$, the set of optimal solutions to (L0-H) is

$$\mathbf{conv}\left\{ \left(\mathscr{L}(\bar{\beta}), \bar{\beta}, z\right) \;\middle|\; \bar{\beta} \textit{ is globally optimal to } (L_0),\ z \in \{0,1\}^{|\mathscr{I}|}, \sum_{j \in \mathscr{I}} z_j \le K \textit{ and } z_j \ge |\bar{\beta}_j|_0, \forall j \in \mathscr{I} \right\}. \tag{6}$$

Proof To prove the algebraic characterization of $\mathbf{clconv}\mathscr{F}(\mathscr{L},K)$, we show that the assumptions in Theorem 1 hold. Recall the union representation (4). Let δ_J be the convex function such that $\delta_J(\beta) = 0$ if $\beta_j = 0$ for all $j \in \mathscr{I} \setminus J$ and $\delta_J(\beta) = +\infty$ otherwise. One can compute that (e.g., by [35, Corollary 8.3.3]),

$$\begin{aligned} \mathscr{R}(C^J) &= \mathscr{R}(\mathbf{epi}(\mathscr{L} + \delta_J) \times \{\mathbf{1}^J\}) = \mathbf{epi}(\mathscr{L}0^+ + \delta_J) \times \{0\} \\ &= \left\{(\tau,\beta,0) \;\middle|\; \mathscr{L}0^+(\beta) \le \tau,\ \beta_j = 0,\ \forall j \in \mathscr{I} \setminus J \right\} \end{aligned}$$

Suppose there exists vectors $(\tau^J,\beta^J,0) \in \mathscr{R}(C^J)$ for all $J \subseteq \mathscr{I}$, $|J| \le K$, and that $\sum_J \tau^J = 0$, $\sum_J \beta^J = 0$. Since $\mathscr{L}$ is assumed to be bounded below, $\tau^J \ge 0$ for all J. Therefore $\tau^J = 0$ for all J. By Assumption 1, $(0,\beta^J,0) \in \mathscr{R}(C^J)$ implies that we must have $\beta^J = 0$ for all J. Then our characterization of $\mathbf{clconv}\mathscr{F}(\mathscr{L},K)$ immediately follows from the characterization of $\mathscr{R}(C^J)$.

Now we characterize the global optimal solution set of (L0-H). By definition and Proposition 1 we have $\eta^* = h^*$. It is straightforward to show that all extreme points in the set in (5) and (6) are optimal to (L0-H) by choosing optimal μ such that it contains only one nonzero entry. So (5) and (6) are subsets of the set of global solutions to (L0-H).

Suppose that $\left(\sum_J \bar{t}^J, \sum_J \bar{\beta}^J, \sum_J \bar{\mu}^J \mathbf{1}^J\right)$ is an optimal solution to (L0-H), we claim that the following inequality holds for all J,

$$\bar{t}^J + \lambda \bar{\mu}^J |J| \ge \bar{\mu}^J h^*. \tag{7}$$

To see this, first consider J such that $\bar{\mu}^J = 0$. We have $\mathscr{L}0^+(\bar{\beta}^J) \le \bar{t}^J$ and $\bar{\beta}^J_j = 0$ for all $j \in \mathscr{I} \setminus J$. By Assumption 1 and the optimality assumption we have $\bar{t}^J = 0$ and $\bar{\beta}^J = 0$, hence (7) holds. For any J such that $\bar{\mu}^J > 0$, $(\bar{\mu}^J)^{-1}\bar{\beta}^J$ is a feasible solution to (L_0) with objective value $\mathscr{L}((\bar{\mu}^J)^{-1}\bar{\beta}^J) + \lambda|J| = (\mu^J)^{-1}\bar{t}^J + \lambda|J| \ge \mu^* = h^*$, which again implies (7). Since $\sum_J \bar{t}^J + \lambda\bar{\mu}^J|J| = h^*$, we have (7) holds as equality for all J. For any J such that $\bar{\mu}^J > 0$ this implies that $(\bar{\mu}^J)^{-1}\bar{t}^J + \lambda|J| = \mathscr{L}((\bar{\mu}^J)^{-1}\bar{\beta}^J) + \lambda|J| = h^*$, i.e., $(\bar{\mu}^J)^{-1}\bar{\beta}$ is a global optimal solution to (L_0). Therefore

$$\left(\sum_J \bar{t}^J, \sum_J \bar{\beta}^J, \sum_J \bar{\mu}^J \mathbf{1}^J\right) = \sum_{J:\bar{\mu}^J > 0} \bar{\mu}^J \left((\bar{\mu}^J)^{-1}\bar{t}^J, (\bar{\mu}^J)^{-1}\bar{\beta}^J, \mathbf{1}^J\right)$$

is the desired convex combination. Suppose $\lambda > 0$, then for all J such that $\bar{\mu}^J > 0$, $\|\bar{\beta}^J\|_0 = \mathbf{1}^J$ otherwise $\mu^J)^{-1}\bar{\beta}^J$ cannot be a global optimal solution to (L_0). □

Proposition 2 *Suppose that Assumption 1 holds and $K \geq 1$, then* $\textbf{clconv}\mathscr{F}(\mathscr{L}, K)$ *is full-dimensional.*

Proof Adding a constant to $\mathscr{L}(\bullet)$ if necessary, we can assume that $\mathscr{L}(0) = 0$. We construct a linearly independent set of $1 + p + |\mathscr{I}|$ nonzero vectors in $\mathscr{F}(\mathscr{L}, K)$. Together with the zero vector, which is also in $\mathscr{F}(\mathscr{L}, K)$, this proves the full-dimensionality. Firstly, let $z_j = \beta_j = 0$ for all $j \in \mathscr{I}$. Since there is no restriction on entries $\{\beta_j\}_{j \notin \mathscr{I}}$ and t can be arbitrarily large, one can find a linearly independent set of $1 + p - |\mathscr{I}|$ vectors in this form. Now for each $j \in \mathscr{I}$, we can find two vectors with $(z_j, \beta_j) = (1, 0)$ and $(z_j, \beta_j) = (1, 1)$ and all other entries in j to be zero. This gives us exactly $1 + p + |\mathscr{I}|$ linearly independent nonzero vectors in $\mathscr{F}(\mathscr{L}, K)$.

2.1 The block separable case

The convex formulation (L0-H) can quickly become impractical as K gets larger. However, if $\mathscr{L}(\beta)$ has a block structure, such exponential growth can be potentially avoided. Consider the special case that

$$\mathscr{L}(\beta) = \sum_{i=1}^{m} \mathscr{L}^{(i)}(\beta^{(i)}). \tag{8}$$

where $\beta = \left[\beta^{(1)}, \beta^{(2)}, ..., \beta^{(m)}\right]$, $\beta^{(i)} \in \mathbb{R}^d$, and $\mathscr{L}^{(i)} : \mathbb{R}^d \mapsto \mathbb{R}$ is convex for $i = 1, ..., m$. Suppose for simplicity that $\mathscr{I} = \{1, 2, ..., md\}$, then a convex relaxation of (L_0) is the following *block-wise convexification*

$$\zeta_{block} = \min \quad \sum_{i=1}^{m} t^{(i)} + \lambda \sum_{i=1}^{m} \sum_{j=1}^{d} z_j^{(i)}, \qquad s.t., \ \ (t^{(i)}, \beta^{(i)}, z^{(i)}) \in \textbf{clconv}\mathscr{F}(\mathscr{L}^{(i)}, d), \ \ \sum_{i=1}^{m} \sum_{j=1}^{d} z_j^{(i)} \leq K. \tag{9}$$

It is easy to see that $\zeta_{block} \leq \zeta_{L0}$. An attractive feature of this convex relaxation is that its problem size grows linearly in m when d is fixed. This block separable structure (8) appears in applications such as the spatial graphical model [18]. Unfortunately (9) is in general not equivalent to the complete convexification (L0-H), unless the cardinality constraint is inactive at some optimal solution.

Proposition 3 *Suppose $\mathscr{L}(\bullet)$ has the block separable form (8), and $\mathscr{I} = \{1, ..., md\}$. If there exists a solution $\left\{\left(\bar{t}^{(i)}, \bar{\beta}^{(i)}, z^{(i)}\right)\right\}_{i=1}^{m}$ optimal to (9) such that $\sum_{i=1}^{m} \sum_{j=1}^{d} \bar{z}_j^{(i)} < K$, then $\zeta_{block} = \zeta_{L0}$.*

Proof If such an optimal solution exists, then it is optimal to (9) without the inequality constraint. Then for each i, $(\bar{t}^{(i)}, \bar{\beta}^{(i)}, z^{(i)})$ is optimal to the following problem

$$\min \quad t^{(i)} + \lambda \sum_{j=1}^{d} z_j^{(i)} \qquad (t^{(i)}, \beta^{(i)}, z^{(i)}) \in \textbf{clconv}\mathscr{F}(\mathscr{L}^{(i)}, d).$$

By Theorem 2 there exists an optimal solution to (L_0) at which the cardinality constraint is inactive. Therefore $\zeta_{L0} = \zeta_{block}$.

Note that this proposition holds when there is no cardinality constraint. The discussion of perspective relaxations and related penalty functions in the next section exploits this insight when $d = 1$.

3 Perspective relaxation by separable decomposition

The previous general convexification (L0-H) can lead to very large convex optimization problems if K in the definition of $\mathscr{F}(\mathscr{L}, K)$ is large. An alternative approach already in use in the area of mixed-integer nonlinear optimization [20,25,26,42] is to decompose $\mathscr{L}(\bullet)$ into the sum of a non-separable component and a separable component, and to convexify the binary indicators functions together with the separable component. In this section we show this partial convexification approach leads to problems equivalent to some nonconvex regularization functions developed in statistical learning community. For simplicity in the rest of our paper we only focus on the following penalized formulation without cardinality constraints, while assuming $\lambda > 0$ and $\mathscr{I} = \{1, ..., p\}$, i.e.,

$$\zeta_{L0} := \min_{\beta} \quad \mathscr{L}(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|_0. \tag{L0-pen}$$

For most results in the rest of our paper, parallel results can be developed for the general formulation (L_0).

Our exploration is partially motivated by the following interpretation that the classical lasso model can also be seen as the continuous relaxation of a mixed-integer convex formulation of (L_0). Consider the following problem with $\lambda > 0$ and M sufficiently large,

$$\min_{\beta, z} \ \mathscr{L}(\beta) + \lambda \sum_{i=1}^{p} z_i, \ \text{ s.t. } |\beta_i| \le M z_i, \ z_i \in \{0,1\}, \ \forall 1 \le i \le p, \tag{MIQP$_{\lambda,M}$}$$

and the popular lasso model [37] ,

$$\min_{\beta} \ \mathscr{L}(\beta) + \bar{\lambda} \sum_{i} |\beta_i|. \tag{lasso}$$

We have the following relation.

Proposition 4 *A continuous relaxation of (MIQP$_{\lambda,M}$), where each binary conditions* $z_i \in \{0,1\}$ *is relaxed to* $z_i \in [0, +\infty)$*, is equivalent to (lasso) with penalty parameter* $\bar{\lambda} = \frac{\lambda}{M}$.

Proof When the binary conditions $z_i \in \{0,1\}, \forall i$ are relaxed to $z_i \in [0, +\infty), \forall i$, as $\lambda > 0$, z_i must take the value $\frac{|\beta_i|}{M}$ in an optimal solution to (MIQP$_{\lambda,M}$). Therefore this continuous relaxation is equivalent to the lasso model with penalty parameter $\bar{\lambda} = \lambda / M$. □

It is then natural to ask does more sophisticated convex relaxation techniques for (L_0) have connections with other penalty functions used in statistical learning. We consider the approach of *perspective relaxation* (see, e.g., [21,25,26]), which can be considered as a special application of the disjunctive formulation discussed in the previous section. We then show its equivalence and relation with the *minimax concave penalty* (MCP) [41] and the *reverse Huber penalty* [34].

3.1 Perspective Relaxation, Minimax Concave Penalty, and Reverse Huber Penalty

The perspective relaxation has been show to be useful in deriving convex relaxations of mixed-integer nonlinear programming with balanced strength and complexity. Given p scalar convex functions $f_i(\bullet):\mathbb{R}\mapsto\mathbb{R}$, we decompose $\mathscr{L}(\bullet)$ as

$$\mathscr{L}(\beta)=\left[\mathscr{L}(\beta)-\sum_i f_i(\beta_i)\right]+\sum_i f_i(\beta_i) \tag{10}$$

and apply our convexification approach to the separable summand. We obtain the following relaxation (not necessarily convex),

$$\begin{aligned}\zeta_{PR}\left(\{f_i\}_{i=1}^p\right):=\min_\beta\quad &\left[\mathscr{L}(\beta)-\sum_i f_i(\beta_i)\right]+\sum_{i=1}^p t_i+\lambda\sum_{i=1}^p z_i\\ s.t.,\quad &(t_i,\beta_i,z_i)\in\mathbf{clconv}\mathscr{F}(f_i,1),\ \forall 1\le i\le p.\end{aligned} \tag{PR}$$

If $\mathscr{L}(\beta)=\sum_{i=1}^p f_i(\beta_i)$, i.e., $\mathscr{L}$ is separable by itself, (PR) is exactly the block-wise convexification (9) when $d=1$. It is an exact convexification of (L0-pen) by Proposition 3. In general the optimal value $\zeta_{PR}\left(\{f_i\}_{i=1}^p\right)$ is always a lower bound of ζ_{L0}, because for any $\bar\beta$ feasible in (L_0), $(\bar t,\bar\beta,\bar z)$ with $\bar t_i=f_i(\bar\beta_i)$ and $z_i=|\bar\beta_i|_0$ $(\forall i)$, is feasible in (PR) with the same objective value. The characterization of $\mathbf{clconv}\mathscr{F}(f,1)$ for a scalar convex function f is well-known in the literature, however we present a specialized version for the sake of completeness. Recall that $\tilde f$ is the perspective function of f as defined in (1).

Proposition 5 *Let $f(\beta):\mathbb{R}\mapsto\mathbb{R}$ be a scalar convex function. Then*

$$\mathbf{clconv}\mathscr{F}(f,1)=\left\{\left((1-z)f(0)+\tilde t,\beta,z\right)\ \middle|\ \tilde f(z,\beta)\le\tilde t,\ z\in[0,1]\right\}.$$

Proof Straightforward by applying Theorem 2 with $J=\emptyset$ and $J=\{1\}$.

A special case that will be useful in our analysis later is when $f_i(\beta_i)=\delta_i\beta_i^2$ where $\delta_i\ge 0$. It can be easily verified (and known in the literature, e.g., [20,21,13,42]) that

$$\mathbf{clconv}\mathscr{F}(\delta_i\beta_i^2,1)=\left\{\left(t_i,\beta_i,z_i\right)\ \middle|\ \delta_i\beta_i^2\le t_iz_i,\ t_i\ge 0,z_i\in[0,1]\right\}. \tag{11}$$

Note that the nonlinear inequality in (11) is a (rotated) second-order-cone constraint.

By projecting out the additional variables t and z, we obtain the following equivalent form of (PR):

Theorem 3 *(PR) is equivalent to the following problem*

$$\min_\beta\ \mathscr{L}(\beta)+\sum_i\rho_{f_i}(\beta_i;\lambda), \tag{$PR:reg$}$$

where

$$\rho_{f_i}(\beta_i;\lambda)=\min_{z_i\in[0,1]}\left\{(1-z)f_i(0)+\tilde f_i(z_i,\beta_i)-f_i(\beta_i)+\lambda z_i\right\} \tag{12}$$

The equivalence is in the sense that β^* is a global optimal solution to $(PR:reg)$ if and only if there exists t^* and z^* such that (t^*,β^*,z^*) is a global optimal solution to (PR).

Proof Observe that the objective function in (PR) is $\mathscr{L}(\beta)+\sum_i \big(t_i - f_i(\beta_i)\big)+\lambda z_i$. It is easy to see (PR) can be reformulated as

$$\min_{\beta} \mathscr{L}(\beta)+\sum_i \rho_{f_i}(\beta_i;\lambda),$$

where

$$\begin{aligned}\rho_{f_i}(\beta_i;\lambda) &= \min_{t_i,z_i}\left\{\big(t_i - f_i(\beta_i)\big)+\lambda z_i \;\middle|\; (t_i,\beta_i,z_i)\in \mathbf{clconv}\mathscr{F}(f_i,1)\right\}\\ &= \min_{z_i\in[0,1]}\left\{(1-z_i)f_i(0)+\tilde{f}_i(z_i,\beta_i)-f_i(\beta_i)+\lambda z_i\right\}.\end{aligned} \tag{13}$$

The last equality is straightforward by Proposition 5. □

We call $\rho_{f_i}(\bullet;\lambda)$ in (12) *perspective penalty functions*. Different choices of f_i may lead to different penalty functions. However, the following observation suggests that what matters is the “nonlinearity" in f_i.

Remark 2 (Invariance under affine addition.) Let $f:\mathbb{R}\mapsto\mathbb{R}$ be a scalar convex function and $\rho_f(\beta;\lambda)$ is the penalty function of the form in (12),

$$\rho_f(\beta;\lambda)=\inf_{z\in[0,1]}\left\{(1-z)f(0)+\tilde{f}(z,\beta)-f(\beta)+\lambda z\right\}.$$

Let $h:\mathbb{R}\mapsto\mathbb{R}$ be an (scalar) affine function, i.e., $h(x)=cx+d$ for some $c,d\in\mathbb{R}$. It is easy to verify that $(\widetilde{f+h})(z,\beta)=\tilde{f}(z,\beta)+c\beta+dz$ and

$$\begin{aligned}\rho_{f+h}(\beta;\lambda) &= (1-z)[f(0)+d]+\tilde{f}(z,\beta)+c\beta+dz-f(\beta)-c\beta-d+\lambda z\\ &= (1-z)f(0)+\tilde{f}(z,\beta)-f(\beta)+\lambda z=\rho_f(\beta;\lambda).\end{aligned}$$

An immediate implication is that it suffices to assume $f_i(0)=0$ in the decomposition (10).

Applied to the quadratic splitting where $f_i(\beta_i)=\delta_i\beta_i^2$ for $\delta_i\geq 0$, $i=1,\ldots,p$. The following proposition provides a closed-form expression for the perspective penalities.

Proposition 6 *Let $f_i(\beta_i)=\delta_i\beta_i^2$ for $\delta_i\geq 0$ and $i=1,\ldots,p$. Then*

$$\rho_{f_i}(\beta_i;\lambda)=\begin{cases}2\sqrt{\delta_i\lambda}|\beta_i|-\delta_i\beta_i^2, & if\ \delta_i\beta_i^2\leq\lambda;\\ \lambda, & if\ \delta_i\beta_i^2>\lambda.\end{cases} \tag{14}$$

Proof Note that $f_i(0)=0$ and

$$\tilde{f}_i(z_i,\beta_i)-f_i(\beta_i)+\lambda z_i=\begin{cases}\delta_i\beta_i^2/z_i-\delta_i\beta_i^2+\lambda z_i, & z_i\in(0,1]\\ 0, & z_i=0,\beta_i=0,\\ +\infty, & z_i=0,\beta_i\neq 0.\end{cases} \tag{15}$$

When $\beta_i=0$, it is easy to see that $\rho_{f_i}(0;\lambda)=0$. Otherwise since $\delta_i\beta_i^2/z_i+\lambda z_i\geq 2\sqrt{\lambda\delta_i\beta_i^2}$ with equality holds when $z_i=\sqrt{\delta_i\beta_i^2/\lambda}$. It is then left to verify that

$$\rho_{f_i}(\beta_i;\lambda)=\min_{z_i\in[0,1]}\left\{\tilde{f}_i(z_i,\beta_i)-f_i(\beta_i)+\lambda z_i\right\}=\begin{cases}2\sqrt{\delta_i\lambda}|\beta_i|-\delta_i\beta_i^2, & \text{if } \delta_i\beta_i^2\leq\lambda;\\ \lambda, & \text{if } \delta_i\beta_i^2>\lambda.\end{cases}$$

□

For notational purpose we will use $\rho_{\delta_i}(\beta_i;\lambda)$ to denote $\rho_{f_i}(\beta_i;\lambda)$ when $f_i(\beta_i) = \delta_i\beta_i^2$ for all i. Several remarks are in order.

Remark 3 (Problem Convexity) The penalty function (14) is nonconvex in β_i. However, (PR) as well as $(PR: reg)$ with $f_i(\beta_i) = \delta_i\beta_i^2$, $\delta_i \geq 0$, can be a convex program when $\mathscr{L}(\beta) - \sum_i \delta_i\beta_i^2$ is convex, i.e., when $\mathscr{L}(\bullet)$ is strongly convex and all δ_i's are sufficiently small (but nonnegative).

Remark 4 (Under-estimation property) Since (PR) is a valid relaxation of (L_0), it is not a surprise that in the equivalent penalization form, $\rho_{f_i}(\beta_i;\lambda) \leq \lambda|\beta_i|_0$ for all β_i and all choices of f_i. Observe that $\tilde{f}_i(z_i,0) = z_i f_i(0)$ and $\tilde{f}_i(1,\beta_i) = f(\beta_i)$, by definition (12),

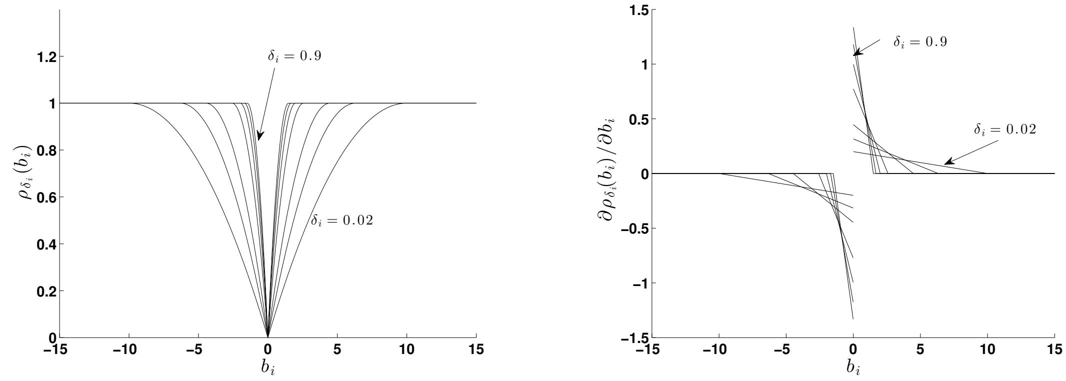
$$\rho_{f_i}(0;\lambda) = \min_{z_i\in[0,1]} \left\{(1-z_i)f_i(0) + z_i f_i(0) - f_i(0) + \lambda z_i\right\} = 0, \quad \forall \lambda \geq 0.$$

Take $z_i = 1$, it is easy to verify that $\rho_{f_i}(\beta_i;\lambda) \leq \lambda$ for all $\beta_i \in \mathbb{R}$.

Remark 5 (Connection with other penalty functions.) In fact, the formula (14) is a rediscovery of the Minimax Concave Penalty (MCP) proposed by Zhang [41]. The MCP is defined as follows (with $\hat{\lambda} \geq 0$ and $\hat{\gamma} > 0$),

$$\rho(t;\hat{\lambda}) := \hat{\lambda}\int_0^{|t|}\left(1 - \frac{x}{\hat{\gamma}\hat{\lambda}}\right)_+ dx = \begin{cases} \hat{\lambda}|t| - t^2/(2\hat{\gamma}) & \text{if } |t| \leq \hat{\gamma}\hat{\lambda}, \\ \hat{\gamma}\hat{\lambda}^2/2 & \text{if } |t| > \hat{\gamma}\hat{\lambda}. \end{cases}$$

By identifying λ and δ_i in (14) with $\hat{\lambda}^2/2$ and $2/\hat{\gamma}$ respectively, it is easy to verify the equivalence of (14) and MCP up to this change of variables. In [41] one single parameter $\tilde{\gamma}$ is used to control the concavity of penalty functions for all i. This corresponds to the special case of $f_i(\beta_i) = \delta_i\beta_i^2$ where all δ_i take the same value. Figure 1 illustrates the penalty function (14) with $\lambda = 1$ and different choices of parameter δ_i. With fixed δ_i, this function is continuously differentiable at any nonzero value, and its second derivative is $-\delta_i$ when $\beta_i \in \left[-\sqrt{\frac{\lambda}{\delta_i}}, 0\right) \cup \left(0, \sqrt{\frac{\lambda}{\delta_i}}\right]$. When β_i is fixed, $\rho_{\delta_i}(\beta_i)$ is a concave function of δ_i when $\delta_i > 0$.



(a) Penalty function

(b) Penalty gradient

Fig. 1: Illustration of penalty function (14)

In [34], a reformulation with convex objective and binary variables is obtained for a cardinality constrained case of (L_0). When specialize to the following ℓ_2 - ℓ_0 regularized problem with $\mu > 0$,

$$\min_{\beta} \underbrace{\frac{1}{2}\|X\beta - y\|_2^2 + \frac{1}{2}\mu\|\beta\|_2^2}_{\mathscr{L}(\beta)} + \lambda\|\beta\|_0, \tag{L2L0}$$

they obtained a convex relaxation with an equivalent penalty form with a *reverse Huber penalty*. We show it can also be derived as a special case of (14). Their equivalent penalty form is [2],

$$\min_{\beta} \frac{1}{2}\|X\beta - y\|_2^2 + 2\lambda\sum_{i=1}^{p} B\left(\sqrt{\frac{\mu}{2\lambda}}\beta_i\right) \tag{16}$$

where B denotes the reverse Huber penalty

$$B(t) = \begin{cases} |t| & \text{if } |t| \le 1, \\ \frac{t^2+1}{2} & \text{otherwise.} \end{cases}$$

Take $f_i(\delta_i) = f_i(\mu/2) = (\mu/2)\beta_i^2$, it is straightforward to verify that $\rho_{f_i}(\beta_i;\lambda)$ in (14) equals

$$\rho_{\delta_i}(\beta_i;\lambda) = -\frac{\mu}{2}\beta_i^2 + 2\lambda B\left(\sqrt{\frac{\mu}{2\lambda}}\beta_i\right).$$

So (16) is equivalent to $(PR: reg)$ with such choices of f_i. Note that the authors of [34] also discovered the second-order-cone representability of the reverse Huber penalty (proof of Corollary 3 in [34]), although not derived from perspective relaxations, i.e., characterization (11).

3.2 The Optimal Convex Perspective Relaxation

It is natural to ask how to choose f_i in the decomposition (10). A general study of this question is beyond the scope of this paper, while we only consider the special case where each $f_i(\beta_i)$ is convex quadratic under a stipulation that $\{f_i\}_{i=1}^p$ are chosen so that $(PR: reg)$ is a *convex* problem. By the invariance of addition of affine functions (Remark 2), it suffices to assume $f_i(\beta_i) = \delta_i\beta_i^2$ where $\delta_i \ge 0$ for all i. Let $\zeta_{PR}(\delta)$ to denote $\zeta_{PR}\left(\{f_i\}_{i=1}^p\right)$ in this case, where δ is the vector with i-th entry δ_i. As previously mentioned, $\zeta_{PR}(\delta)$ is always a lower bound of ζ_{L0}, and usually computable if $\mathscr{L}(\beta) - \sum_i \delta_i\beta_i^2$ is convex. This is possible for some positive δ if and only if $\mathscr{L}$ is strongly convex, and we assume in this section and the rest of the paper:

Assumption 2 *For all $i = 1,\dots,p$, $f_i(\beta_i) = \delta\beta_i^2$ for some $\delta_i \ge 0$.*

Assumption 3 *$\mathscr{L}: \mathbb{R}^p \mapsto \mathbb{R}$ is strongly convex.*

[2] The difference of a constant factor from Corollary 3 in [34] is due to a typo in their derivation.

We further define set C as,

$$C := \left\{ \delta \in \mathbb{R}_+^p \,\middle|\, \mathscr{L}(\beta) - \sum_i \delta_i \beta_i^2 \quad \text{is convex} \right\}.$$

We aim to find the best parameter $\delta \in C$ such that $\zeta_{PR}(\delta)$ is as large as possible. This question is formulated as the following max-min problem.

$$\max_{\delta \in C} \underbrace{\min_{\beta} \left\{ \mathscr{L}(\beta) + \sum_i \rho_{\delta_i}(\beta_i; \lambda) \right\}}_{:=\zeta_{PR}(\delta)}. \tag{Max-Min}$$

A closely related formulation by interchanging the “min" and “max" operators has the interpretation of simultaneously exploiting infinitely many penalty functions corresponding to all $\delta \in C$,

$$\min_{\beta} \max_{\delta \in C} \left\{ \mathscr{L}(\beta) + \sum_i \rho_{\delta_i}(\beta_i; \lambda) \right\}. \tag{Min-Max}$$

We study this minimax pair and its computable representation in this section. Before our technical development, we remark that this notion of “optimal perspective relaxation" is not new, and has appeared in a previous work by some of the authors [13] as well as [42] in a context of bounded feasible regions. Results in those works do not immediately apply to our context because (1) they restrict to the case of convex quadratic objective functions and (2) they assume boundedness of variables. Furthermore, our minimax approach presented here is more intuitive and transparent than the derivation based on conic duality as in [13,42].

Since C is compact, the classical minimax theory [35, Corollary 37.3.2] guarantees that the objective value of these two problems are equal. Additional conditions are needed to prove a saddle point exists. We use the following theorem in convex analysis.

Theorem 4 (Theorem 37.6 in [35]) *Let $K(\delta, b)$ be a closed proper concave-convex function with effective domain $C \times D$. If both of the following conditions,*

1. *The convex functions $K(\delta, \cdot)$ for $\delta \in \mathbf{ri}(C)$ have no common direction of recession;*
2. *The convex functions $-K(\cdot, b)$ for $b \in \mathbf{ri}(D)$ have no common direction of recession;*

are satisfied, then K has a saddle-point in $C \times D$. In other words, there exists $(\delta^*, b^*) \in C \times D$, such that

$$\inf_{b \in D} \sup_{\delta \in C} K(\delta, b) = \sup_{\delta \in C} \inf_{b \in D} K(\delta, b) = K(\delta^*, b^*).$$

The following theorem applies Theorem 4 in our context.

Theorem 5 *Suppose that Assumption 2 and 3 hold. Let $\zeta_{\max\min}$ and $\zeta_{\min\max}$ be the optimal values of (Max-Min) and (Min-Max) respectively. There exists $\delta^* \in C$ and $\beta^* \in \mathbb{R}^p$ such that*

$$\zeta_{\max\min} = \mathscr{L}(\beta^*) + \sum_i \rho_{\delta_i^*}(\beta_i^*; \lambda) = \zeta_{\min\max}.$$

Proof Let function $K(\delta,\beta)$ be defined as follows

$$K(\delta,b) := \begin{cases} \mathscr{L}(\beta)+\sum_i \rho_{\delta_i}(\beta_i;\lambda), & \forall \delta \in C, \\ -\infty, & \forall \delta \notin C. \end{cases} \tag{17}$$

It is easy to check that $K(\delta,b)$ is concave in δ and convex in β, a so-called *concave-convex function*. The *effective domain* of $K(\cdot,\cdot)$ is defined as

$$\{\delta \mid K(\delta,\beta) > -\infty, \forall \beta\} \times \{\beta \mid K(\delta,\beta) < +\infty, \forall \delta\} = C \times \mathbb{R}^p,$$

which is nonempty. So $K(\bullet,\bullet)$ is *proper*. Further K is closed as for each fixed β, the function $K(\cdot,\beta)$ is upper-semicontinuous, and for each fixed $\delta \in C$, $K(\delta,\cdot)$ is lower-semicontinuous. Finally, for any $\beta \in \mathbb{R}^p$, $-K(\cdot,b)$ has no direction of recession as C is bounded. For any $\delta \in \mathbf{ri}(C)$, $K(\beta,\bullet)$ is also strongly convex, therefore has no direction of recession. Then our conclusion directly follows from Theorem 4. □

An algorithm for computing the saddle point could be potentially designed if it is tractable to check the convexity of $\mathscr{L}(\beta)-\sum_i \delta_i \beta_i^2$ for fixed δ. However this is difficult in general even for quartic polynomials [1]. In the rest of this paper we suppose that $\mathscr{L}(\beta)$ is (strongly) convex and quadratic, i.e.,

Assumption 4 *$\mathscr{L}(\bullet)$ is (strongly) convex and quadratic, i.e. $\mathscr{L}(\beta) = 0.5\beta^T Q\beta - c^T\beta$, where $Q \succ 0$ and $c \in \mathbb{R}^p$.*

In this case the set $C = \{\delta \in \mathbb{R}^p_+ \big| Q - \mathbf{D}(\delta) \succeq 0\}$. It has been proved in different settings [13,42] that "best perspective relaxations" with convex quadratic objective functions and binary indicators is equivalent to some semidefinite relaxations. However, these existing results do not directly apply to our setting (L0-pen) due to the unboundedness of β. We provide a direct proof to verify the equivalence between a semidefinite relaxation and the minimax pair (Max-Min) – (Min-Max) with guarantees of attainment of minimizers.

We propose the following semidefinite relaxation for (L_0),

$$\zeta_{SDP} := \min_{\beta,z,B} \; 0.5\langle Q,B\rangle - c^T\beta + \lambda\sum_i z_i, \;\text{ s.t. }\; B \succeq \beta\beta^T, \;\begin{bmatrix} z_i & \beta_i \\ \beta_i & B_{ii}\end{bmatrix} \succeq 0, \forall i. \tag{SDP}$$

It is easy to check that this is a valid relaxation to (L0-pen) with Assumption 4. Indeed, if the convex constraint $B \succeq \beta\beta^T$ were replaced by $B = \beta\beta^T$, then (at optimality) $z_i = 1$ if $\beta_i \neq 0$ and $z_i = 0$ if $\beta_i = 0$. Some basic properties of (SDP) is summarized in the following proposition.

Proposition 7 *Suppose that Assumption 2 and 4 hold and $\lambda > 0$. Let (β^*, z^*, B^*) be an optimal solution to (SDP), then for all i, $z_i^* \in [0,1]$. Further, $z_i^* = \frac{(b_i^*)^2}{B_{ii}^*}$ if $B_{ii}^* \neq 0$, and value $z_i^* = 0$ otherwise. If B^* is a rank-1 matrix, then z_i^* is binary for all i, and β^* is an (global) optimal solution to (L0-pen).*

Proof As z_i only appears in the objective and the constraint $\begin{bmatrix} z_i & \beta_i \\ \beta_i & B_{ii}\end{bmatrix} \succeq 0$, if (β^*, z^*, B^*) is an optimal solution to (SDP), then $z_i = \frac{\beta_i^2}{B_{ii}}$ if $B_{ii} > 0$ and 0 otherwise. Note that $B_{ii} \geq \beta_i^2$ is implied by the constraint $B \succeq bb^T$, $z_i^* \in [0,1]$.

If B^* is rank-1, by $B^* \succeq \beta^* \left(\beta^*\right)^T$ we have $B^* = \beta^* \left(\beta^*\right)^T$. Therefore in this case $z_i^* = 1$ if $B_{ii}^* \neq 0$, and $z_i^* = 0$ otherwise. It is then easy to see that

$$0.5\langle Q, B^*\rangle - c^T\beta^* + \lambda\sum_i z_i^* = \mathscr{L}(\beta^*) + \lambda\sum_i |\beta_i^*|_0,$$

Since (SDP) is a relaxation of (L0-pen), β^* is optimal to (L0-pen). □

Note that (SDP) is meaningful only when $Q \succ 0$ (i.e., Assumption 3 holds). Otherwise, if Q has a nontrivial null space, e.g., $\exists d \neq 0, X^T X d = 0$, then by following the recession direction $B \mapsto B + \tau d d^T$, as $\tau \mapsto +\infty$, B_{ii} may become arbitrarily large and $z_i \mapsto 0$ for all i such that $d_i \neq 0$.

The dual problem to (SDP) is

$$\begin{aligned} \zeta_{SDP} = \sup_{\epsilon\in\mathbb{R},\alpha\in\mathbb{R}^p,\delta,t\in\mathbb{R}^p} \quad & -\epsilon/2 \\ \text{s.t.} \quad & \begin{bmatrix} \epsilon & \alpha^T \\ \alpha & Q - \mathbf{D}(\delta) \end{bmatrix} \succeq 0, \quad \begin{bmatrix} \delta_i & -c_i - \alpha_i \\ -c_i - \alpha_i & 2\lambda \end{bmatrix} \succeq 0, \forall i. \end{aligned} \tag{DSDP}$$

It is easy to see that both (SDP) and (DSDP) are both strictly feasible if $\lambda > 0$ and $Q \succ 0$. Therefore strong duality holds and the optimal value is attained at some primal optimal solution (β^*, z^*, B^*) and dual optimal solution $(\epsilon^*, \alpha^*, \delta^*)$. The following theorem shows the equivalence of primal-dual pair (SDP) and (DSDP) with the minimax pair (Min-Max) and (Max-Min).

Theorem 6 *Suppose Assumption 2 and 4 holds and* $\lambda > 0$. *Let* (β^*, z^*, B^*) *and* $(\epsilon^*, \alpha^*, \delta^*)$ *be optimal solutions to (SDP) and (DSDP), respectively, then* (δ^*, β^*) *is a saddle point for (Min-Max) and (Max-Min).*

Proof Let $K(\cdot,\cdot)$ and C be defined as in (17), (β^*, z^*, B^*) and $(\epsilon^*, \alpha^*, \delta^*)$ be optimal solutions to (SDP) and (DSDP) respectively. We would like to show that for all $\delta \in C$, $b \in \mathbb{R}^p$,

$$\max_{\delta\in C} K(\delta, \beta^*) = \zeta_{SDP} = \min_{\beta\in\mathbb{R}^p} K(\delta^*, \beta). \tag{18}$$

Provided (18), (δ^*, β^*) is a saddle point because

$$K(\delta^*, \beta^*) \le \max_{\delta\in C} K(\delta, \beta^*) = \min_{b\in\mathbb{R}^p} K(\delta^*, \beta) \le K(\delta^*, \beta^*).$$

As both inequalities hold as equalities, (δ^*, β^*) is the desired saddle point, i.e.,

$$K(\delta, \beta^*) \le K(\delta^*, \beta^*) \le K(\delta^*, \beta), \quad \forall \delta \in C, \beta \in \mathbb{R}^p.$$

By our derivation of (PR) and $(PR: reg)$, $\zeta_{PR}(\delta) = \min_\beta K(\delta, \beta) \le K(\delta, \beta^*)$. By taking maximization over $\delta \in C$ on both sides, we obtain $\max_{\delta\in C} \zeta_{PR}(\delta) \le \max_{\delta\in C} K(\delta, \beta^*)$. Together with the equality $\zeta_{PR}(\delta^*) = \min_\beta K(\delta^*, \beta)$, to prove (18), it suffices to show

$$\max_{\delta\in C} \zeta_{PR}(\delta) = \zeta_{SDP} = \zeta_{PR}(\delta^*).$$

Firstly, we show that for any $\delta \in C$, $\zeta_{SDP} \ge \zeta_{PR}(\delta)$. In fact, let $(\bar{\beta}, \bar{z}, \bar{B})$ be an optimal solution to (SDP). We construct a feasible solution to (PR) with objective value no larger than ζ_{SDP}. By equation (11) and

Proposition 7, $(\delta_i \bar{B}_{ii}, \bar{\beta}_i, \bar{z}_i) \in \textbf{clconv}\mathscr{F}(\delta_i \beta_i^2, 1)$. So $(\delta \circ \mathbf{D}(\bar{B}), \bar{\beta}, \bar{z})$ is feasible in (PR). To compare the objective values, we have

$$\begin{aligned} 0.5\langle Q, \bar{B}\rangle - c^T\bar{\beta} + \lambda \sum_i \bar{z}_i &= 0.5\langle Q, \bar{\beta}\bar{\beta}^T\rangle - c^T\bar{\beta} + \lambda \sum_i \bar{z}_i + 0.5\langle Q, \bar{B} - \bar{\beta}\bar{\beta}^T\rangle \\ &\geq 0.5\langle Q, \bar{\beta}\bar{\beta}^T\rangle - c^T\bar{\beta} + \lambda \sum_i \bar{z}_i + 0.5\langle \mathbf{D}(\delta), \bar{B} - \bar{\beta}\bar{\beta}^T\rangle \\ &= 0.5\bar{\beta}^T(Q - \mathbf{D}(\delta))\bar{\beta} - c^T\bar{\beta} + 0.5\sum_i \delta_i \bar{B}_{ii} + \lambda \sum_i \bar{z}_i. \end{aligned}$$

The inequality is due to the fact that $Q - \mathbf{D}(\delta) \succeq 0$ and $\bar{B} - \bar{\beta}\bar{\beta}^T \succeq 0$. Therefore we have $\max_{\delta\in C} \zeta_{PR}(\delta) \leq \zeta_{SDP}$. Now we show $\zeta_{SDP} \leq \zeta_{PR}(\delta^*)$, which will then complete the proof of (18) by

$$\zeta_{PR(\delta^*)} \leq \max_{\delta\in C} \zeta_{PR(\delta)} \leq \zeta_{SDP} \leq \zeta_{PR(\delta^*)}.$$

We achieve this by showing the optimal value of (DSDP) is less than or equal to the objective value of any feasible solution to (PR) with $f_i(\beta_i) = \delta_i^* \beta_i^2$. Let $(\bar{t}, \bar{\beta}, \bar{z})$ denote such a feasible solution, we have two sets of matrix inequalities

$$\begin{bmatrix} \epsilon^* & \alpha^{*T} \\ \alpha^* & Q - \mathbf{D}(\delta^*) \end{bmatrix} \succeq 0, \quad \begin{bmatrix} 1 & \bar{\beta}^T \\ \bar{\beta} & \bar{\beta}\bar{\beta}^T \end{bmatrix} \succeq 0, \tag{19}$$

$$\begin{bmatrix} \delta_i^* & -c_i - \alpha_i^* \\ -c_i - \alpha_i^* & 2\lambda \end{bmatrix} \succeq 0, \quad \begin{bmatrix} \bar{t}_i & \bar{\beta}_i \\ \bar{\beta}_i & \bar{z}_i \end{bmatrix} \succeq 0, \quad \forall i. \tag{20}$$

As the inner product between two matrices in (19) is nonnegative, we have

$$\zeta_{SDP} = -\frac{1}{2}\epsilon^* \leq \alpha^{*T}\bar{\beta} + \frac{1}{2}\bar{\beta}\left(Q - \mathbf{D}(\delta^*)\right)\bar{\beta}.$$

Next by taking the inner product between the matrices in (20) we obtain

$$\alpha^{*T}\bar{\beta} \leq -\sum_i \bar{\beta}_i c_i + \frac{1}{2}\sum_i \delta_i^* \bar{t}_i + \sum_i \lambda \bar{z}_i.$$

Therefore,

$$\zeta_{SDP} \leq \frac{1}{2}\bar{\beta}^T(Q - \mathbf{D}(\delta^*))\bar{\beta} - c^T\bar{\beta} + \frac{1}{2}\sum_i \delta_i^* \bar{t}_i + \sum_i \lambda \bar{z}_i.$$

Since $(\bar{t}, \bar{\beta}, \bar{z})$ is an arbitrarily chosen feasible solution, $\zeta_{SDP} \leq \zeta_{PR(\delta^*)}$. This completes our proof. □

4 Randomized Rounding by the Goemans-Williamson Procedure

Proposition 7 states that if the optimal B in (SDP) is rank-1, then it certifies a global optimal solution to (L0-pen). When the optimal B is not rank-1, it is reasonable to ask how to randomly round an optimal solution to (SDP) to obtain good feasible solutions to (L0-pen). Celebrated examples of rounding SDP solutions to obtain high-quality solutions to the original combinatorial problem include max-cut [24] and binary quadratic programs [33], where bounds on the relative accuracy of the semidefinite relaxation was obtained. We show that by reformulating (L_0) as a two-level problem, where the inner problem is a binary quadratic program, and (SDP) is equivalent to replacing the inner problem with its natural semidefinite relaxation.

Recall that ζ_{L0} is the optimal value of (L0-pen) with Assumption 4, and ζ_{SDP} is the optimal value of (SDP). ζ_{L0} can be written as

$$\zeta_{L0} = \min_{u\in\mathbb{R}^p} \ \zeta_{BQP}(u). \tag{21}$$

where

$$\zeta_{BQP}(u) := \min_{z\in\{0,1\}^p} \ \frac{1}{2} z^T \left[\mathbf{D}(u) Q \mathbf{D}(u)\right] z - (c\circ u)^T z + \lambda \sum_{i=1}^{p} z_i. \tag{BQP(u)}$$

A well-known semidefinite relaxation for (BQP(u)) in the primal form is

$$\begin{aligned} \zeta_{MCSDP}(u) = \min_{z,Z} \ & \frac{1}{2} \langle \mathbf{D}(u) Q \mathbf{D}(u), Z\rangle + (\lambda e - c\circ u)^T z \\ s.t., \quad & Z \succeq zz^T, Z_{ii} = z_i, \forall i = 1,\ldots,p. \end{aligned} \tag{MCSDP(u)}$$

Note that quadratic programming with binary variables can be formulated as a max cut problem, which typically use variables encoded by $\{-1,1\}$ in the literature. It is straightforward to verify that by the following change of variable,

$$t \longleftarrow T(z) := \begin{bmatrix} 1 & 0^T \\ -e & 2I \end{bmatrix} \begin{bmatrix} 1 \\ z \end{bmatrix},$$

(MCSDP(u)) is equivalent to the semidefinite relaxation for the max cut problem discussed in the literature, e.g., [24,33]. In fact, we have the following equivalence in parallel of (21).

Theorem 7 *Suppose $\lambda > 0$ in (L0-pen) and Assumption 4 holds, we have*

$$\zeta_{SDP} = \min_{u\in\mathbb{R}^p} \ \zeta_{MCSDP}(u). \tag{22}$$

Let (β^*, z^*, B^*) be an optimal solution to (SDP). Define u^* and Z^* as follows

$$u_i^* := \begin{cases} B_{ii}^*/\beta_i^*, & if \ \beta_i^* \neq 0 \\ 0, & if \ \beta_i^* = 0 \end{cases} \quad \forall i, \quad and \quad Z_{ij}^* = \begin{cases} \frac{B_{ij}^* \beta_i^* \beta_j^*}{B_{ii}^* B_{jj}^*}, & if B_{ii}^* B_{jj}^* \neq 0, \\ 0, & if B_{ii}^* B_{jj}^* = 0, \end{cases} \quad \forall i,j, \tag{23}$$

then (z^*, Z^*) is optimal to $MCSDP(u^*)$. If $u^* \in \arg\min_{u\in\mathbb{R}^p} \zeta_{MCSDP}(u)$ with (z^*, Z^*) optimal to MCSDP(u^*), then $(u^*\circ z^*, z^*, \mathbf{D}(u^*) Z^* \mathbf{D}(u^*))$ is optimal to (SDP). The value 0 in the definition of u_i^* can be replaced with arbitrary value in $\mathbb{R}$.

Proof It is straightforward to verify that for any $(\bar{z}, \bar{Z})$ is feasible to MCSDP(u), then $(\bar{\beta}\circ\bar{z}, \bar{z}, \mathbf{D}(\bar{u}) Z \mathbf{D}(\bar{u}))$ is feasible to (SDP) with the same objective value, because of identity $\langle Q, \mathbf{D}(\bar{u}) Z \mathbf{D}(\bar{u})\rangle = \langle \mathbf{D}(\bar{u}) Q \mathbf{D}(\bar{u}), Z\rangle$. This implies that $\zeta_{SDP} \le \zeta_{MCSDP}(u)$ for all $u \in \mathbb{R}^p$.

On the other hand, suppose (β^*, B^*, z^*) is optimal to (SDP), u^* and Z^* are defined as stated. We show that (z^*, Z^*) is feasible to MCSDP(u^*) with the same objective value, which proves that $\zeta_{SDP} \ge \zeta_{MCSDP}(u^*)$.

Indeed, for all j such that $B^*_{jj} = 0$, by Proposition 7 $z^*_j = 0$ and by definition $Z^*_{ij} = Z^*_{ji} = 0$ $(\forall i)$. Let J to denote the set $\left\{ j \,\middle|\, B^*_{jj} \neq 0 \right\}$, then for all $i, j \in J$, we have

$$\left(Z^* - z^* \left(z^* \right)^T \right)_{ij} = \frac{B^*_{ij} \beta^*_i \beta^*_j}{B^*_{ii} B^*_{jj}} - \frac{\left(\beta^*_i \beta^*_j \right)^2}{B^*_{ii} B^*_{jj}} = \left(B^*_{ij} - \beta^*_i \beta^*_j \right) \frac{B^*_{ij} \beta^*_i \beta^*_j}{B^*_{ii} B^*_{jj}}.$$

Therefore $Z^* - z^* (z^*)^T$ is the Hadamard product of two positive semidefinite matrices and therefore positive semidefinite. Finally for all $j \in J$, again by Proposition 7 and definition of Z^*, $Z^*_{ii} = (\beta^*_i)^2 / B^*_{ii} = z^*_i$. It is left to see that

$$0.5\langle Q, B^* \rangle - c^T \beta^* + \lambda \sum_j z^*_j = 0.5\langle \mathbf{D}(u^*) Q \mathbf{D}(u^*), Z^* \rangle + \lambda (e - c \circ u^*)^T z^*,$$

which can be proved by checking $B^*_{ij} = Z^*_{ij} u^*_i u^*_j$ for all i, j and $\beta^*_i = u^*_i z_i$ for all i. This concludes our proof of (22). The correspondence of optimal solutions is clear by our discussion. □

Motivated by the relation (21) and Theorem 7, given a solution (β^*, z^*, B^*) optimal to (SDP), we may interpret it as an optimal solution to (MCSDP(u)) with $u = u^*$ (where u^* is defined as in Theorem 7). A binary solution to (BQP(u)) with $u = u^*$ can be constructed by randomized rounding, and used to reconstruct an approximate solution to (L0-pen). A detailed description is given in Algorithm 1. Note that in implementation, closed-form formulae can be used to solve the subproblems in step 7.

We conclude this section with two remarks. One reviewer pointed out that in [34], a cardinality constrained case of (L_0) is reformulated as a convex program with binary variables by exploiting a two-level formulation that essentially switches the order of inner and outer problems in (21). See their proof of Theorem 1 in [34]. For the case of (strongly) convex quadratic loss $\mathscr{L}(\beta)$, their approach leads to a convex relaxation with the reverse Huber penalty (16). It is interesting that changing the order leads to different relaxations, and deeper connections between these two approaches are worth exploring in future work. For example, whether their binary convex formulation can lead to insights in choosing the decomposition (10) for general loss functions.

Another reviewer asked about whether one could derive approximation guarantees of (SDP). The main results in [33] imply that for any *fixed* u, $|\zeta_{MCSDP}(u) - \zeta_{BQP}(u)|$ can be bounded by a constant fraction of $|\zeta^+_{BQP}(u) - \zeta_{BQP}(u)|$, where $\zeta^+_{BQP}(u)$ is the objective value of the *worst* binary solution to (BQP(u)). However, it does not seem easy to obtain a nontrivial bound on $|\zeta_{SDP} - \zeta_{L0}|$ by using these results. The difficulty seems to lie in controlling the quality of optimal u^* in equation (22).

5 Numerical Results

In this final section we provide some numerical experiments to compare our semidefinite relaxation (SDP), the convex relaxation proposed in [34], and the performance of branch-and-bound algorithm (implemented in Gurobi) with a time limit.

We consider the following problem,

$$\min_b \; \frac{1}{2n} \|X\beta - y\|_2^2 + \frac{1}{2}\mu \|\beta\|_2^2 + \lambda \|\beta\|_0. \tag{24}$$

Algorithm 1: A randomized rounding algorithm for (SDP)

```
Data: An optimal solution to (SDP), denoted by $(\beta^*, z^*, B^*)$, and parameter $N$ (number of random points in the
        Goemans-Williamson rounding procedure);
Result: An approximate solution $\hat{\beta}$ to (L0-pen);
Generate matrix $Z^* \in \mathscr{S}^p$ by (23);
Generate matrix $T^* \in \mathscr{S}^{p+1}$ by
        $$T^* := \begin{bmatrix} 1 & 0^T \\ -e & 2I \end{bmatrix} \begin{bmatrix} 1 & (z^*)^T \\ z^* & Z^* \end{bmatrix} \begin{bmatrix} 1 & -e^T \\ 0 & 2I \end{bmatrix};$$
Compute a factorization $T^* = U^* (U^*)^T$, where $U^* \in \mathbb{R}^{p \times r}$;
Randomly generate vectors $\left\{v^{(1)}, ..., v^{(N)}\right\} \subseteq \mathbb{R}^r$ from the normal distribution with mean 0 and covariance matrix $I$;
For each $k = 1, ..., N$, compute $t^{(k)} \leftarrow \mathbf{sign}(U^* r)$; If $t_1^{(k)} = -1$, $t^{(k)} \leftarrow -t^{(k)}$;
For each $k = 1, ..., N$, compute vector $z^{(k)} \in \{0,1\}^p$ by
        $$z_j^{(k)} \leftarrow 0.5\left(t_{j+1}^{(k)} + 1\right), j = 1, ..., p;$$
For each $k = 1, ..., N$, compute $v^{(k)}$ by solving the following convex quadratic program in a restricted subspace
        $$v^{(k)} = \lambda \sum_j z^{(k)} + \min_{\beta \in \mathbb{R}^p} \left\{0.5\beta^T Q\beta - c^T\beta \;\middle|\; \beta_j = 0\ \forall j \text{ s.t., } z_j^{(k)} = 0\right\};$$
let $\beta^{(k)}$ denote an optimal solution to this problem;
Let $K$ be the index such that $v^{(K)}$ is the smallest in $\left\{v^k\right\}_k$. Then the output vector is set as $\hat{\beta} := \beta^{(K)}$.
```

where $X \in \mathbb{R}^{n \times p}$ with each row being an observation of predictor variables, and $y \in \mathbb{R}^n$ is the vector of observed responses. This is in the form of (L0-pen) with convex quadratic loss function (**??**) where $Q = \frac{1}{2n}X^T X + \frac{1}{2}\mu I$ and $c = \frac{1}{n}X^T y$. In all instances considered below, each entry of X is i.i.d. Gaussian $\mathcal{N}(0,1)$, and each column of X is rescaled to have standard deviation 1 – so that all diagonal entries in $\frac{1}{n}X^T X$ are 1. Vector y is generated based on an assumed underlying truth β^{true}, which is simulated by

$$\beta_i^{true} = \begin{cases} U_{[-1,1]}, & i = 1, ..., k, \\ 0, & i = k+1, ..., p, \end{cases}$$

where $U_{[-1,1]}$ is the uniform distribution $[-1,1]$. Then y is generated by

$$y = X\beta^{true} + \epsilon, \quad \text{where } \epsilon_i \sim \mathcal{N}\left(0, \left[\sigma(X\beta^{true})/SNR\right]^2\right),\ \forall i.$$

$\sigma(X\beta^{true})$ is the standard deviation of the "signal vector" $X\beta^{true}$. Parameter SNR (signal-noise ratio) is used to control the noise level in a transparent way, i.e., $SNR = \sigma(X\beta^{true})/\sigma(\epsilon)$. The following MIQP formulation is used to solve (24) with Gurobi, where M is set as $5\|\beta^{true}\|_\infty$,

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2n}\|X\beta - y\|_2^2 + \frac{1}{2}\mu\|\beta\|_2^2 + \lambda \sum_j z_j,\ \ \beta_j \in [-Mz_j, Mz_j],\ \ z_j \in \{0,1\},\ \ \forall j = 1, ..., p.$$

We generate data with $n > p$ as well as $n < p$, $SNR \in \{1,3\}$ and different values of λ and μ. For each pair of parameters (λ, μ), we randomly generate 5 instances[3]. For each instance, the time limit for Gurobi is

[3] Our first version of the paper contains 30 instances for a subset of parameter settings considered here. However we observe the variation of performance among instances within a same parameter setting is small.

set as 60 seconds, and default options (as set by Yamip [31]) are used (with “Method=-1 (automatic)"). We denote the best upper and lower bounds obtained by Gurobi in the time limit as τ_{UB} and τ_{Grb}. The optimal values of convex relaxation (16), as well as (SDP) applied to (24), are computed and denoted by τ_{PWG} and τ_{SDP}, respectively. Then three kinds of relative gap are computed by

$$\text{GrbGap} = \frac{\tau_{UB} - \tau_{Grb}}{\tau_{UB}} \times 100\%, \;\; \text{SDPGap} = \frac{\tau_{UB} - \tau_{SDP}}{\tau_{UB}} \times 100\%, \;\; \text{PWGGap} = \frac{\tau_{UB} - \tau_{PWG}}{\tau_{UB}} \times 100\%.$$

All experiments are run on a workstation with Interl Core i5, which has a max clock speed 2.7GHz.

As discussed in previous sections, the PWG bound is a special case of perspective relaxation, and theoretically cannot be tighter than the optimal perspective relaxation bound (SDP). This is confirmed in our computational results where (SDP) is always tighter than the PWG bound, although the degree of improvements vary according to parameter settings. We also observe that the upper and lower bounds of Gurobi stabilize well in the 60-second time limit, and progresses slowly afterwards. Since for all instances (SDP) is solved in 10-40 seconds, it is reasonable to compare such bounds.

Three sets of test instances are considered with different difficulty levels. The “Easy" instances have parameters $(n,p,k) = (100,60,10)$ and moderate noise level ($SNR = 3$). Gurobi solves all instances to optimality in less than 60 seconds with thousands of nodes. In all choices of λ and μ, SDPGap is always significantly smaller than PWGGap, and often nearly 0 for problems with smaller λ values (hence the problem is in a sense less “discrete"). The “Medium" instances have $(n,p,k) = (100,60,10)$ and higher noise level ($SNR = 1$). For these instances SDPGap is always smaller than 1/3 of PWGGap. SDPGap is also comparable to GrbGap (and can be better with smaller λ values). $(n,p,k) = (60,100,10)$ and $SNR = 3$ in the “Hard" instances. For those instances GrbGap is quite large. PWGGap is always better, especially when λ is small or μ is larger. The improvement of SDPGap upon PWGGap is in the range of 20% to 50%.

Table 1: Average relative gap of Gurobi, (SDP) and convex relaxation in [34]

		“Easy"			“Medium"			“Hard"		
		$(n,p,k) = (100,60,10), SNR = 3$			$(n,p,k) = (100,60,10), SNR = 1$			$(n,p,k) = (60,100,10), SNR = 3$		
	Avg. Gap	$\mu = 0.1$	$\mu = 0.2$	$\mu = 0.3$	$\mu = 0.1$	$\mu = 0.2$	$\mu = 0.3$	$\mu = 0.1$	$\mu = 0.2$	$\mu = 0.3$
$\lambda = 0.05$	SDP	0.49%	0.02%	0.02%	1.56%	1.35%	0.65%	8.98%	3.48%	3.15%
	PWG	3.97%	1.24%	0.69%	7.32%	4.70%	2.57%	10.65%	4.58%	4.29%
	Gurobi	0.00%	0.00%	0.00%	3.92%	6.00%	3.96%	28.43%	20.67%	26.13%
	(#nodes)	(1.3E3)	(1.3E03)	(1.5E03)	(2.8E5)	(2.6E5)	(2.7E5)	(7.4E5)	(7.6E05)	(7.8E05)
$\lambda = 0.10$	SDP	0.01%	0.11%	0.00%	3.18%	1.43%	1.06%	8.81%	3.14%	2.26%
	PWG	9.21%	3.34%	1.55%	11.25%	5.95%	3.96%	12.03%	5.33%	4.00%
	Gurobi	0.00%	0.00%	0.00%	4.84%	3.75%	4.78%	18.52%	16.40%	18.45%
	(#nodes)	(3.6E2)	(6.4E2)	(1.5E3)	(2.6E5)	(2.6E5)	(2.3E5)	(5.9E5)	(7.3E5)	(7.6E5)
$\lambda = 0.20$	SDP	2.69%	1.69%	2.16%	4.83%	2.27%	1.53%	20.33%	9.88%	7.27%
	PWG	22.74%	12.70%	10.48%	17.21%	8.36%	6.51%	24.81%	13.73%	10.23%
	Gurobi	0.00%	0.00%	0.00%	3.60%	1.67%	1.84%	24.80%	20.97%	19.02%
	(#nodes)	(1.3E3)	(2.8E3)	(6.0E3)	(2.3E5)	(2.5E5)	(2.5E5)	(5.4E5)	(6.5E5)	(7.4E5)

We now consider the effectiveness of Goemans-Williamson rounding in our context. We applied Algorithm 1 with 1000 random vectors to (SDP) solutions on all instances considered in the previous experiment. Note that Algorithm 1 is very efficient. For all instances we consider here, Algorithm 1 completes in a small fraction of second. In implementation redundant vectors in $\{z^{(k)}\}$ obtained step 6 of Algorithm

1 can be ignored, and subproblems in step 7 can be solved in closed-form (where only a small linear system needs to be solved for each k). Let τ_{GW} denote the best objective value of problem (L2L0) found by our rounding procedure. In majority of cases we have $\tau_{GW} \geq \tau_{UB}$, i.e., the upper bounds obtained by the rounding procedure are no better than those found by Gurobi in the time limit of 60 seconds. However $\tau_{GW} = \tau_{UB}$ for 45/45 “Easy" instances, 36/45 “Medium" instances and 28/45 “Hard" instances. We only find 2 instances where $\tau_{GW} < \tau_{UB}$. This suggests that all upper bounds found by Gurobi have very high quality. In Table 2, we report the averaged relative differences

$$\frac{\tau_{GW} - \tau_{UB}}{\tau_{UB}} \times 100\%$$

for each pair of choices of λ and μ. These results show that the rounding procedure can also find high quality solutions, with the exception of “Hard" instances with large λ and small μ (so the problem is in a sense most “discrete"). When $\lambda = 0.2$ and $\mu = 0.1$, the best upper bound found by the rounding procedure is 6% worse than the upper bound found by Gurobi.

Table 2: Averaged relative difference of upper bounds by Algorithm 1 and Gurobi

	“Easy"			“Medium"			“Hard"		
	(n,p,k)=(100,60,10), SNR=3			(n,p,k)=(100,60,10), SNR=1			(n,p,k)=(60,100,10), SNR=3		
	$\mu = 0.1$	$\mu = 0.2$	$\mu = 0.3$	$\mu = 0.1$	$\mu = 0.2$	$\mu = 0.3$	$\mu = 0.1$	$\mu = 0.2$	$\mu = 0.3$
$\lambda = 0.05$	0.00%	0.00%	0.00%	0.02%	0.01%	0.03%	0.10%	0.00%	0.03%
$\lambda = 0.10$	0.00%	0.00%	0.00%	0.28%	0.09%	0.03%	2.34%	0.00%	0.00%
$\lambda = 0.20$	0.00%	0.00%	0.00%	0.00%	0.10%	0.00%	6.55%	0.78%	0.60%

We finally comment on the computational cost of solving (SDP). The size of (SDP) is primarily determined by p – the number of predictor variables in regression, while does not depend on n. Also note that (SDP) has a relatively “clean" form, i.e., the number of linear constraints is small, and in fact grows linearly with respect to p. The dual-scaling interior point algorithm for SDP [4] is especially suitable for solving such SDP problems to high accuracy. In table 3 we report the typical computational time needed to solve one instance of (SDP) as p increases in table, using the software DSDP [3] implemented by Benson, Ye and Zhang, with their default parameters.

Table 3: Computational time (seconds) to solve (SDP) with DSDP [3]

$p = 50$	$p = 100$	$p = 200$	$p = 400$	$p = 800$
0.33	1.20	4.58	37.48	278.9

These results show that (SDP) can be actually a practical way to certify the quality of feasible solutions to (L0-pen) with (strongly) convex quadratic loss function. In context where (SDP) need to be solved many times for different choices of λ, it makes sense to consider cheaper approximate algorithms, such as the first-order algorithms, that also benefit from warm-starting when λ is slightly changed. An especially attractive approach is to use low rank factorizations and nonlinear programming [11]. We will leave comprehensive computational studies for future work.

6 Conclusions

Our paper is an attempt of connecting some techniques in mixed-integer nonlinear optimization and variable selection with sparsity-inducing penalty functions. The main contribution of this paper lies in applying convexification techniques to (L_0) and demonstrate their relations to existing techniques in statistical learning. We derived the complete (disjunctive) convexification formulation of (L_0), perspective penalty functions depending on a decomposition of the loss function (10), and for the case of strongly convex loss functions, an optimal perspective relaxation. The perspective penalty function generalizes multiple penalties including the minimax penalty function (MCP), which is often among the best penalties in various studies, and the reverse Huber penalty in [34]. The optimal perspective relaxation can be used to certify the qualities of “local" or approximate solutions to (L_0), and improves upon some existing techniques for this purpose.

In statistical learning it is often argued that good “local" solutions to highly nonconvex problems may be as good as the global optimum from a statistical perspective. However in practice it is difficult to certify the quality of “local" feasible solutions. Polynomial time computable lower bounds on the optimal value, which are natural products of convexification techniques, can be helpful in closing this gap from the computational side. Some open questions left in this paper include (1) how to choose the decomposition (10) according to different classes of loss functions, for example polyhedral loss functions; (2) how to strengthen (SDP), and to combine with other lower bounding techniques, to obtain stronger lower bounds in order to certify the quality of feasible solutions to (L_0), etc.

Acknowledgement. The authors would like thank two anonymous reviewers for comments and suggestions that lead to significantly improvements in various aspects of this paper.

References

1. Ahmadi, A.A., Olshevsky, A., Parrilo, P.A., Tsitsiklis, J.N.: Np-hardness of deciding convexity of quartic polynomials and related problems. Mathematical Programming **137**(1–2), 453–476 (2013)
2. Albert, A., Anderson, J.A.: On the existence of maximum likelihood estimates in logistic regression models. Biometrika **71**(1), 1–10 (1984)
3. Benson, S.J., Ye, Y.: DSDP5: Software for semidefinite programming. Tech. Rep. ANL/MCS-P1289-0905, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL (2005). URL `http://www.mcs.anl.gov/~benson/dsdp`. Submitted to ACM Transactions on Mathematical Software
4. Benson, S.J., Ye, Y., Zhang, X.: Solving large-scale sparse semidefinite programs for combinatorial optimization. SIAM Journal on Optimization **10**(2), 443–461 (2000)
5. Bertsimas, D., King, A., Mazumder, R.: Best subset selection via a modern optimization lens. submitted to Annals of Statistics (2014)
6. Bertsimas, D., Shioda, R.: Algorithm for cardinality-constrained quadratic optimization. Computational Optimization and Applications **43**(1), 1–22 (2009)
7. Bickel, P.J., Ritov, Y., Tsybakov, A.B.: Simultaneous analysis of Lasso and Dantzig Selector. The Annals of Statistics **37**(4), 1705–1732 (2009)
8. Bienstock, D.: Computational study of a family of mixed-integer quadratic programming problems. Mathematical Programming, Series A **74**(2), 121–140 (1996)
9. Breheny, P., Huang, J.: Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Annals of Applied Statistics **5**(1), 232–253 (2011)
10. Bühlmann, P., van de Geer, S.: Statistics for High-Dimensional Data. Springer (2009)
11. Burer, S., Monteiro, R.: A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. Mathematical Programming (Series B) **95**, 329–357 (2003)
12. Ceria, S., Soares, J.: Convex programming for disjunctive convex optimization. Math. Program., Ser. A **86**(3), 595–614 (1999)

13. Dong, H., Linderoth, J.: On valid inequalities for quadratic programming with continuous variables and binary indicators. In: IPCO 2013: The Sixteenth Conference on Integer Programming and Combinatorial Optimization, vol. 7801, pp. 169–180. Springer (2013)
14. Efron, B., Hastie, T.J., Johnstones, I., Tibshirani, R.J.: Least angle regression. Annals of Statistics **32(2)**, 407–499 (2004)
15. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association **96**(456), 1348–1360 (2001). DOI 10.2307/3085904
16. Fan, J., Lv, J.: A selective overview of variable selection in high dimensional feature space. Statist. Sinica **20**(1), 101–148 (2010)
17. Fan, J., Lv, J.: Nonconcave penalized likelihood with np-dimensionality. IEEE Trans. Inf. Theor. **57**(8), 5467–5484 (2011). DOI 10.1109/TIT.2011.2158486
18. Fang, E.X., Liu, H., Wang, M.: Blessing of massive scale: Spatial graphical model estimation with a total cardinality constraint. Under review. Available at `http://www.optimization-online.org/DB_FILE/2015/11/5205` (2015)
19. Feng, M., Mitchell, J.E., Pang, J.S., Shen, X., W achter, A.: Complementarity formulations of ℓ_0-norm optimization problems. Journal submission
20. Frangioni, A., Gentile, C.: Perspective cuts for a class of convex 0-1 mixed integer programs. Mathematical Programming **106**, 225–236 (2006)
21. Frangioni, A., Gentile, C.: SDP diagonalizations and perspective cuts for a class of nonseparable MIQP. Operations Research Letters **35**(2), 181–185 (2007)
22. Frank, I.E., Friedman, J.H.: A Statistical View of Some Chemometrics Regression Tools. Technometrics **35**(2), 109–135 (1993). DOI 10.2307/1269656. URL `http://www.jstor.org/stable/1269656`
23. Friedman, J., Hastie, T.J., Höfling, H., Tibshirani, R.: Pathwise coordinate optimization. Annals of Applied Statistics **2**, 302–332 (2007)
24. Goemans, M., Williamson, D.: .878-approximation algorithms for MAX CUT and MAX 2SAT. Proceedings of the Symposium of Theoretical Computer Science pp. 422–431 (1994)
25. Günlük, O., Linderoth, J.: Perspective reformulations of mixed integer nonlinear programming with indicator variables. Mathematical Programming (Series B) **124**(1-2), 183–205 (2010)
26. Günlük, O., Linderoth, J.T.: Perspective reformulation and applications. In: J. Lee, S. Leyffer (eds.) The IMA Volumes in Mathematics and its Applications, vol. 154, pp. 61–92 (2012)
27. Huang, J., Breheny, P., Ma, S.: A selective review of group selection in high dimensional models. Statist. Sci. **27**(4), 481–499 (2012)
28. Huang, J., Horowitz, J.L., Ma, S.: Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Annals of Statistics **36**(2), 587–613 (2008)
29. Huang, J., Ma, S., Zhang, C.H.: Adaptive lasso for high-dimensional regression models. Statistica Sinica **18**, 1603–1618 (2008)
30. Lange, K.: Optimization. Springer Texts in Statistics. Springer-Verlag, New York (2004)
31. Löfberg, J.: Yalmip: A toolbox for modeling and optimization in matalb. In: Proceedings of the CACSD Conference. Taipei, Taiwan (2004)
32. Lv, J., Fan, Y.: A unified approach to model selection and sparse recovery using regularized least squares. Annals of Statistics **37**(6A), 3498–3528 (2009). DOI 10.1214/09-aos683
33. Nesterov, Y.: Quality of semidefinite relaxation for nonconvex quadratic optimization. CORE discussion paper 9719, Center for Operations Research & Econometrics (1997)
34. Pilanci, M., Wainwright, M.J., Ghaoui, L.E.: Sparse learning via Boolean relaxations. Mathematical Programming (Series B) **151**, 63–87 (2015)
35. Rockafellar, R.T.: Convex Analysis. Princeton University Press (1970)
36. Stubbs, R.A., Mehrotra, S.: A branch-and-cut method for 0-1 mixed convex programming. Mathematical Programming, Series A **86**, 515–532 (1996)
37. Tibshirani, R.J.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B **58**, 267–288 (1996)
38. Tseng, P.: Coordinate ascent for maximizing nondifferentiable concave functions. Technical Report LIDS-P p. 1840 (1988)
39. Wang, H., Leng, C.: Unified lasso estimation by least squares approximation. Journal of the American Statistical Association **102**(479), pp. 1039–1048 (2007). URL `http://www.jstor.org/stable/27639944`
40. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B **68**, 49–67 (2006)
41. Zhang, C.H.: Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics **38**(2), 894–942 (2010)
42. Zheng, X., Sun, X., Li, D.: Improving the performance of miqp solvers for Quadratic Programs with Cardinality and Minimum Threshold Constraints: A Semidefinite Program. INFORMS Journal on Computing **26**(4), 690–7–3 (2014)
43. Zou, H.: The adaptive lasso and its oracle properties. Journal of the American Statistical Association **101**, 1418–1429 (2006)
44. Zou, H., Hastie, T.J.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B **67**(2), 301–320 (2005). DOI 10.1111/j.1467-9868.2005.00503.x
45. Zou, H., Li, R.: One-step sparse estimates in nonconcave penalized likelihood models. Annals of Statistics **36**, 1509–1533 (2008)
46. Zou, H., Zhang, H.H.: On the adaptive elastic-net with a diverging number of parameters. Annals of Statistics **37**, 1733–1751 (2009)