ON THE CONVEXIFICATION OF CONSTRAINED QUADRATIC OPTIMIZATION PROBLEMS WITH INDICATOR VARIABLES

Linchuan Wei
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL, USA
LinchuanWei2022@u.northwestern.edu

Andrés Gómez *
Daniel J. Epstein Department of Industrial and Systems Engineering
University of Southern California
Los Angeles, CA, USA
gomezand@usc.edu

Simge Küçükyavuz†
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL, USA
simge@northwestern.edu

January 14, 2020

ABSTRACT

Motivated by modern regression applications, in this paper, we study the convexification of quadratic optimization problems with indicator variables and combinatorial constraints on the indicators. Unlike most of the previous work on convexification of sparse regression problems, we simultaneously consider the nonlinear objective, indicator variables, and combinatorial constraints. We prove that for a separable quadratic objective function, the perspective reformulation is ideal independent from the constraints of the problem. In contrast, while rank-one relaxations cannot be strengthened by exploiting information from k-sparsity constraint for $k \geq 2$, they can be improved for other constraints arising in inference problems with hierarchical structure or multi-collinearity.

Keywords Convexification · Perspective formulation · Indicator variables · Quadratic optimization · Combinatorial constraints.

1 Introduction

Given a data matrix $X = [x_1, \ldots, x_p] \in \mathbb{R}^{n \times p}$ of features and a response vector $y \in \mathbb{R}^n$, we study constrained regression problems of the form

$$\min_{z,\beta} \|y - X\beta\|_2^2 + \lambda f(\beta) \tag{1a}$$

$$\text{subject to } \beta_i(1 - z_i) = 0 \qquad i \in [p] \tag{1b}$$

$$\beta \in \mathbb{R}^p,\ z \in Q \subseteq \{0,1\}^p, \tag{1c}$$

*Andrés Gómez is supported, in part, by grant 1930582 of the National Science Foundation.
†Simge Küçükyavuz is supported, in part, by ONR grant N00014-19-1-2321.

where β is a vector of regression coefficients, z is a vector of indicator variables with $z_i = 1$ if $\beta_i \neq 0$ (through indicator constraint (1b)), the set Q in constraints (1c) encodes combinatorial constraints on the indicator variables and $[p] = \{1, 2 \ldots, p\}$. The objective (1a) is to minimize the squared loss function plus a regularization term $\lambda f(\beta)$. Typical choices of f include L0, L1 or L2 regularizations.

If Q is defined via a k-sparsity constraint, $Q = \{z \in \{0,1\}^p \mid \sum_{i=1}^p z_i \leq k\}$, then problem (1) reduces to the best subset selection problem [36], a fundamental problem in statistics. Nonetheless, constraints other than the cardinality constraint arise in several statistical problems. Bertsimas and King [9] suggest imposing constraints of the form $\sum_{i \in S} z_i \leq 1$ for some $S \subseteq [p]$ to prevent multicollinearity. Constraints of the form $z_i \leq z_j$ can be used to impose hierarchy constraints [11]. In group variable selection, indicator variables of regression coefficients of variables in the same group are linked, see [32]. Manzour et al. [35] impose that the indicator variables, which correspond to edges in an underlying graph, do not define cycles – a necessary constraint for inference problems with causal graphs. Cozad et al. [16] suggest imposing a variety of constraints in both the continuous and discrete variables to enforce priors from human experts.

Problem (1) is $\mathcal{NP}$-hard [38], and is often approximated with a convex surrogate such as lasso [29, 40]. Solutions with better statistical properties than lasso can be obtained from non-convex continuous approximations [22, 46]. Alternatively, it is possible to solve (1) to optimality via branch-and-bound methods [10, 15]. In all cases, most of the approaches for (1) have focused on the k-sparsity constraint (or its Lagrangian relaxation). For example, a standard technique to improve the relaxations of (1) revolves around the use of the *perspective reformulation* [1, 14, 19, 20, 23, 24, 25, 26, 28, 31, 44, 47], an ideal formulation of a separable quadratic function with indicators (but no additional constraints). Recent work on obtaining ideal formulations for non-separable quadratic functions [3, 4, 5, 20, 27, 33] also ignores additional constraints in Q.

There is a recent research thrust on studying constrained versions of (1). Dong et al. [18] study problem (1) from a continuous optimization perspective (after projecting out the discrete variables), see also [17]. Hazimeh and Mazumder [30] give specialized algorithms for the natural convex relaxation of (1) where Q is defined via hierarchy constraints. Several results exist concerning the convexification of nonlinear optimization problems with constraints [2, 7, 12, 13, 34, 37, 39, 41, 42, 43], but such methods in general do not deliver ideal, compact or closed-form formulations for the specific case of problem (1) with structured feasible regions. In a recent work closely related to the setting considered here, Xie and Deng [45] proves that the perspective formulation is *ideal* if the objective is separable and Q is defined with a k-sparsity constraint. In a similar vein, Bacci et al. [6] show that the perspective reformulation is tight with unit commitment constraints. However, similar results for more general (non-separable) objective functions or constraints are currently not known.

Our contributions and outline. In this paper, we provide a first study (from a convexification perspective) of the interplay between convex quadratic objectives and combinatorial constraints on the indicator variables. Specifically, we generalize the result in Xie and Deng [45] to arbitrary constraints on z. We also show that the rank-one strengthening given in [4] is ideal for k-sparsity with $k \geq 2$. However, we show that the rank-one strengthening can be improved if $k = 1$, or for hierarchy constraints [11, 30]. We conclude our work with a preliminary numerical study on problems with hierarchy constraints showing that the resulting formulations achieve strong relaxations with only a modest increase in the computational effort required to solve the resulting convex formulations.

Notation. Throughout the paper, we adopt the convention that for $a \in \mathbb{R}$, $\frac{a^2}{0} = +\infty$ if $a \neq 0$ and $\frac{a^2}{0} = 0$ when $a = 0$. We let $\mathbf{1}$ be the vector of all ones, and let e_i denote the ith unit vector of appropriate dimension with 1 in ith component and 0's elsewhere. For a set Q, we denote by $\text{conv}(Q)$ its convex hull and by $\text{cl conv}(Q)$ the closure of its convex hull.

2 Convex Hull Results

We present our convex hull results first for separable quadratic functions, followed by the non-separable case.

2.1 Separable Quadratic Function

Consider the mixed-integer epigraph of a separable quadratic function with arbitrary constraints, $z \in Q$, on the indicator variables:

$$W = \left\{ (z, \beta, t) \in Q \subseteq \{0,1\}^p \times \mathbb{R}^p \times \mathbb{R} \mid \sum_{i \in [p]} \beta_i^2 \leq t,\ \beta_i(1 - z_i) = 0\ \forall i \in [p] \right\}.$$

As Theorem 1 below shows, ideal formulations of W can be obtained by applying the perspective reformulation on the separable quadratic term and, *independently*, strengthening the continuous relaxation of Q. This generalizes the result of Xie and Deng [45] for $Q = \{z \in \{0,1\}^p \mid \sum_{i=1}^p z_i \leq k\}$ and the result of Bacci et al. [6] for unit commitment.

Let

$$Y = \left\{ (z, \beta, t) \in \mathbb{R}^{2p+1} \mid \sum_{i \in [p]} \frac{\beta_i^2}{z_i} \leq t, \quad z \in \operatorname{conv}(Q) \right\}.$$

Theorem 1. *Y is the closure of the convex hull of W:* $\operatorname{cl}\operatorname{conv}(W) = Y$.

Proof. Note that inequality $\frac{\beta_i^2}{z_i} \leq t_i$ is precisely the perspective reformulation [24] of a single quadratic term $t_i = \beta_i^2$, thus the validity of the corresponding inequality in Y follows immediately. For any $(a, b, c) \in \mathbb{R}^{2p+1}$ consider the following two problems

$$\min \quad a^\top z + b^\top \beta + ct \qquad \text{subject to} \quad (z, \beta, t) \in W, \tag{2}$$

and

$$\min \quad a^\top z + b^\top \beta + ct \qquad \text{subject to} \quad (z, \beta, t) \in Y. \tag{3}$$

It suffices to show that (2) and (3) are equivalent, i.e., there exists an optimal solution of (3) that is optimal for (2) with the same objective value. If $c = 0, b = \mathbf{0}$, then both (2) and (3) are equivalent to $\min_{z \in Q} a^\top z$. If either $c = 0$ and $b \neq \mathbf{0}$, or $c < 0$, then (2) and (3) are unbounded. When $c > 0$, without loss of generality, we may assume that $c = 1$ by scaling. For any $(z_i, \beta_i) \in [0,1] \times \mathbb{R}$, $i \in [p]$

$$\max_{\alpha_i \in \mathbb{R}} -\alpha_i \beta_i - \frac{\alpha_i^2}{4} z_i = \begin{cases} \frac{\beta_i^2}{z_i} & \text{if } z_i \neq 0, \\ 0 & \text{if } z_i = \beta_i = 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{4}$$

Identity (4) can be proven by taking derivatives with respect to α_i and setting to 0, see also [8]. Hence, for any $\alpha \in \mathbb{R}^p$

$$-\alpha^\top \beta - \sum_{i \in [p]} \frac{\alpha_i^2}{4} z_i \leq \sum_{i \in [p]} \frac{\beta_i^2}{z_i}. \tag{5}$$

In particular, consider the relaxation of (3) obtained by replacing the constraint that $\sum_{i \in [p]} \frac{\beta_i^2}{z_i} \leq t$ with $-b^\top \beta - \sum_{i \in [p]} \frac{b_i^2}{4} z_i \leq t$ (where we let $\alpha = b$ in (5)), i.e.,

$$\min \quad a^\top z + b^\top \beta + t \tag{6a}$$

$$\text{subject to} \quad -b^\top \beta - \sum_{i \in [p]} \frac{b_i^2}{4} z_i \leq t \tag{6b}$$

$$z \in \operatorname{conv}(Q). \tag{6c}$$

Due to constraint (6b), problem (6) is equivalent to

$$\min \quad a^\top z - \sum_{i \in [p]} \frac{b_i^2}{4} z_i \qquad \text{subject to} \quad z \in \operatorname{conv}(Q).$$

Since (6) is equivalent to a linear program (LP) over an integral polyhedron, it must have an integral optimal solution $z^* \in Q$. Let β^* be such that

$$\beta_i^* = \begin{cases} 0 & \text{if } z_i^* = 0, \\ -\frac{b_i}{2} & \text{if } z_i^* = 1. \end{cases}$$

Now if we let $t^* = \sum_{i \in [p]} (\beta_i^*)^2$, then $(z^*, \beta^*, t^*) \in W$ and $b^\top \beta^* + t^* = \sum_{i \in [p]} -\frac{b_i^2}{4} z_i^*$. Thus the optimal values of (2) and (6) coincide. And since (6) is also a relaxation of (3), the optimal values of (2) and (3) coincide. □

2.2 Rank-One Quadratic Function

In this section, we study the epigraph of a (non-separable) rank-one quadratic function with constraints

$$Z_Q = \{(z, \beta, t) \in Q \times \mathbb{R}^p \times \mathbb{R} \mid (\mathbf{1}^\top \beta)^2 \leq t, \beta_i(1 - z_i) = 0, \forall i \in [p]\}$$

for some $Q \subseteq \{0,1\}^p$. We note that ideal formulations for the unconstrained case $Z_{\{0,1\}^p}$ were provided in [4]:

Proposition 1 (Atamtürk and Gómez [4])**.** *The closure of the convex hull of $Z_{\{0,1\}^p}$ is*

$$\text{cl conv}(Z_{\{0,1\}^p}) = \left\{(z, \beta, t) \in [0,1]^p \times \mathbb{R}^{p+1} | (\mathbf{1}^\top \beta)^2 \leq t, \frac{(\mathbf{1}^\top \beta)^2}{\sum_{i\in[p]} z_i} \leq t\right\}.$$

2.2.1 k-sparsity constraint

We first study sets defined by the k-sparsity constraint,

$$Q_1 = \left\{z \in \{0,1\}^p : \sum_{i\in[p]} z_i \leq k\right\},$$

and prove that, under mild conditions, a generalization of the result of Xie and Deng [45] also holds in this case, that is, ideal formulations are achieved by focusing only on the nonlinear objective and indicator constraints.

Theorem 2. *If $k \geq 2$ and integer, then*

$$\text{cl conv}(Z_{Q_1}) = \left\{(z, \beta, t) \in [0,1]^p \times \mathbb{R}^{p+1} \mid (\mathbf{1}^\top \beta)^2 \leq t, \frac{(\mathbf{1}^\top \beta)^2}{\sum_{i\in[p]} z_i} \leq t, \sum_{i\in[p]} z_i \leq k\right\}.$$

Proof. First, note that the validity of the new inequality defining $\text{cl conv}\,(Z_{Q_1})$ follows from Proposition 1. For $a, b \in \mathbb{R}^p$ and $c \in \mathbb{R}$, let's consider the following two optimization problems:

$$\min \quad a^\top z + b^\top \beta + ct \qquad \text{subject to} \quad (z, \beta, t) \in Z_{Q_1}. \tag{7}$$

and

$$\begin{aligned} \min \quad & a^\top z + b^\top \beta + ct && \text{(8a)}\\ \text{subject to} \quad & (\mathbf{1}^\top \beta)^2 \leq t && \text{(8b)}\\ & \frac{(\mathbf{1}^\top \beta)^2}{\sum_{i\in[p]} z_i} \leq t && \text{(8c)}\\ & \sum_{i\in[p]} z_i \leq k && \text{(8d)}\\ & z \in [0,1]^p. && \text{(8e)} \end{aligned}$$

The analysis for cases where $c = 0$ and $c < 0$ is similar to the proof of Theorem 1, and we can proceed with assuming $c = 1$ and $b \in \mathbb{R}^p$. First suppose that b is not a multiple of all-ones vector, then $\exists b_i < b_j$ for some $i, j \in [p], i \neq j$. Let $\bar{z} = e_i + e_j$, $\bar{\beta} = \tau(e_i - e_j)$ for some scalar τ, and $\bar{t} = 0$. Note that $(\bar{z}, \bar{\beta}, \bar{t})$ is feasible for both (7) and (8), and if we let τ go to infinity the objective value goes to minus infinity. So (7) and (8) are unbounded.

Now suppose that $b = \kappa \mathbf{1}^\top$ for some $\kappa \in \mathbb{R}$ and $c = 1$; in this case both (7) and (8) have finite optimal value. It suffices to show that there exists an optimal solution (z^*, β^*, t^*) of (8) that is integral in z^*. If $\sum_{i\in[p]} z_i^* = 0$, then we know $z_i^* = \beta_i^* = 0, \forall i \in [p]$ for both (7) and (8), and we are done. If $0 < \sum_{i\in[p]} z_i^* < 1$ and the corresponding optimal objective value is 0 (or positive), then by letting $z^* = \mathbf{0}$, $\beta^* = \mathbf{0}$ and $t^* = 0$, we get a feasible solution with the same objective value (or better). If $0 < \sum_{i\in[p]} z_i^* < 1$ and (z^*, β^*, t^*) attains a negative objective value, then let $\gamma = \frac{1}{\sum_{i\in[p]} z_i^*}$: $(\gamma z^*, \gamma\beta^*, \gamma t^*)$ is also a feasible solution of (8) with a strictly smaller objective value, which is a contradiction.

Finally, let's consider the case where $\sum_{i\in[p]} z_i^* \geq 1$. In this case, the constraint $(\mathbf{1}^\top \beta)^2 \leq t$ is active and the optimal value is attained when $\mathbf{1}^\top \beta^* = -\frac{\kappa}{2}$ and $t^* = (\mathbf{1}^\top \beta^*)^2$, and (8) has the same optimal value as the LP:

$$\min \quad a^\top z - \frac{\kappa^2}{4} \qquad \text{subject to} \quad 1 \leq \sum_{i\in[p]} z_i \leq k, z \in [0,1]^p.$$

The constraint set of this LP is an interval matrix, so the LP has an integral optimal solution, z^*, hence, so does (8). □

The assumption that $k \geq 2$ in Theorem 2 is necessary. As we show next, if $k = 1$, then it is possible to strengthen the formulation with a valid inequality that uses the information from the cardinality constraint, which was not possible for $k > 1$. Note that the case $k = 1$ is also of practical interest, as set Q_1 with $k = 1$ arises for example when preventing multi-collinearity, see [9].

Proposition 2. *If $k = 1$, then the following inequality is valid for Z_{Q_1}*

$$\sum_{i\in[p]} \frac{\beta_i^2}{z_i} \leq t. \tag{9}$$

Proof. If $k = 1$, then for any $(z, \beta, t) \in Z_{Q_1}$, if $\sum_{i\in[p]} z_i = 0$, then $z_i = \beta_i = 0, \forall i \in [p]$. Hence, by our convention $0 = \sum_{i\in[p]} \frac{\beta_i^2}{z_i} = (\mathbf{1}^\top\beta)^2 \leq t$. Otherwise, $z_j = 1$ for some $j \in [p]$, and $z_i = \beta_i = 0, \forall i \in [p], i \neq j$. Hence $\sum_{i\in[p]} \frac{\beta_i^2}{z_i} = \beta_j^2 = (\mathbf{1}^\top\beta)^2 \leq t$. □

Observe that inequality (9) is not valid if $k \geq 2$, as for example $(\beta_i + \beta_j)^2 < \beta_i^2 + \beta_j^2 \leq \frac{\beta_i^2}{z_i} + \frac{\beta_j^2}{z_j}$ whenever $\beta_i\beta_j < 0$. As we now show, the addition of (9) leads to an ideal formulation of Z_{Q_1} if $k = 1$.

Theorem 3. *If $k = 1$, then*

$$\text{cl conv}(Z_{Q_1}) = \left\{ (z, \beta, t) \in [0, 1]^p \times \mathbb{R}^{p+1} \mid \sum_{i\in[p]} \frac{\beta_i^2}{z_i} \leq t, \sum_{i\in[p]} z_i \leq 1 \right\}.$$

Proof. First, let's consider another mixed integer epigraph:

$$W_{Q_1} = \left\{ (z, \beta, t) \in \{0, 1\}^p \times \mathbb{R}^p \times \mathbb{R} \mid \sum_{i\in[p]} \beta_i^2 \leq t,\ \beta_i(1 - z_i) = 0\ \forall i \in [p],\ \sum_{i\in[p]} z_i \leq 1 \right\}.$$

For $\forall(z, \beta, t) \in Z_{Q_1}$, there exists at most one β_i $i \in [p]$ such that $\beta_i \neq 0$. Hence, $(\mathbf{1}^\top\beta)^2 = \sum_{i\in[p]} \beta_i^2$ and the result follows from Theorem 1. □

2.2.2 Hierarchy constraints

We now consider the hierarchy constraints given by

$$Q_2 = \{z \in \{0, 1\}^p \mid z_p \leq z_i,\ \forall i \in [p - 1]\}.$$

In other words, if $z \in Q_2$ with $z_p = 1$, then $z_i = 1$ for all $i \in [p]$. First, we give a valid inequality for the set Z_{Q_2}, and then show that it is sufficient to describe $\text{cl conv}(Z_{Q_2})$, when added to the continuous relaxation of the original formulation.

Proposition 3. *The following inequality is valid for Z_{Q_2}*

$$\frac{(\mathbf{1}^\top\beta)^2}{\sum_{i\in[p-1]} z_i - (p - 2)z_p} \leq t.$$

Proof. For any $(z, \beta, t) \in Z_{Q_2}$, if $z_p = 1$, then $z_i = 1, \forall i \in [p]$. In this case, $\frac{(\mathbf{1}^\top\beta)^2}{\sum_{i\in[p-1]} z_i - (p-2)z_p} = (\mathbf{1}^\top\beta)^2 \leq t$. If $z_p = 0$, then $\frac{(\mathbf{1}^\top\beta)^2}{\sum_{i\in[p-1]} z_i - (p-2)z_p} = \frac{(\mathbf{1}^\top\beta)^2}{\sum_{i\in[p]} z_i} \leq t$. If $z_i = 0, \forall i$, then $\beta = \mathbf{0}$, and by our convention $0 = \frac{(\mathbf{1}^\top\beta)^2}{\sum_{i\in[p]} z_i} = (\mathbf{1}^\top\beta)^2 \leq t$. If $z_i = 1$ for some $i \in [p]$, then $\frac{(\mathbf{1}^\top\beta)^2}{\sum_{i\in[p]} z_i} \leq (\mathbf{1}^\top\beta)^2 \leq t$. □

To establish the convex hull of Z_{Q_2}, we first give a lemma whose proof is in the Appendix.

Lemma 1. *The extreme points of the polyhedron*

$$Q_g = \left\{ z \in [0, 1]^p \mid \sum_{i\in[p-1]} z_i - (p - 2)z_p \geq 1,\ z_p \leq z_i\ \forall i \in [p - 1] \right\}$$

are integral.

Now we are ready to give an ideal formulation for Z_{Q_2}.

Theorem 4. *The closure of the convex hull of Z_{Q_2} is given by*

$$\operatorname{cl}\operatorname{conv}(Z_{Q_2}) = \Big\{(z,\beta,t) \in [0,1]^p \times \mathbb{R}^{p+1} \mid (\mathbf{1}^\top\beta)^2 \le t,\ z_p \le z_i, \forall i \in [p-1], \\ \frac{(\mathbf{1}^\top\beta)^2}{\sum_{i\in[p-1]} z_i - (p-2)z_p} \le t\Big\}.$$

Proof. For $a, b \in \mathbb{R}^p$ and $c \in \mathbb{R}$, consider the optimization problems:

$$\min \quad a^\top z + b^\top\beta + ct \qquad \text{subject to} \quad (z,\beta,t) \in Z_{Q_2}. \tag{10}$$

and

$$\begin{aligned} \min \quad & a^\top z + b^\top\beta + ct && \text{(11a)}\\ \text{subject to} \quad & (\mathbf{1}^\top\beta)^2 \le t && \text{(11b)}\\ & \frac{(\mathbf{1}^\top\beta)^2}{\sum_{i\in[p-1]} z_i - (p-2)z_p} \le t && \text{(11c)}\\ & z_p \le z_i, \quad \forall i \in [p-1] && \text{(11d)}\\ & z \in [0,1]^p. && \text{(11e)} \end{aligned}$$

Following similar arguments to those in the beginning of the proof of Theorem 2 (with the exception of letting $\bar{z} = \mathbf{1}$ in the corresponding case), we can assume that $c = 1$ and $b = \kappa\mathbf{1}^\top$ for some $\kappa \in \mathbb{R}$; in this case, (10) and (11) have finite optimal value. Suppose (z^*, β^*, t^*) is an optimal solution of (11), then it suffices to show that (z^*, β^*, t^*) is integral in z^*. If $0 = \sum_{i\in[p-1]} z_i^* - (p-2)z_p^* = z_1^* + \sum_{i=2}^{p-1}(z_i^* - z_p^*)$, then by the constraint $z_i \ge z_p$ and non-negativity of z_i's we must have $z_1^* = 0$ and $z_i^* = z_p^*$ for $i = 2, \dots, p-1$. Furthermore, since $z_1^* \ge z_p^*$, we find that $z_p^* = 0$ and $z_i^* = 0, \forall i \in [p-1]$.

If $0 < \sum_{i\in[p-1]} z_i^* - (p-2)z_p^* < 1$ and the corresponding optimal objective value is 0 (or positive), then by letting $z^* = \mathbf{0}$, $\beta^* = \mathbf{0}$ and $t^* = 0$, we obtain a feasible solution with the same (or better) objective value and integral in z^*. Now suppose (z^*, β^*, t^*) attains a negative objective value in (11); let $\gamma = \frac{1}{\sum_{i\in[p-1]} z_i^* - (p-2)z_p^*} > 1$, then $(\gamma z^*, \gamma\beta^*, \gamma t^*)$ is also a feasible solution of (11) because $\gamma^2(\mathbf{1}^\top\beta^*)^2 = \gamma\frac{(\mathbf{1}^\top\beta^*)^2}{\sum_{i\in[p-1]} z_i^* - (p-2)z_p^*} \le \gamma t^*$, for each $i \in [p-1]$ we have $\gamma z_i^* = \frac{z_i^*}{z_i^* + \sum_{j\neq i, j\in[p-1]}(z_j^* - z_p^*)} \le 1$, and $\gamma z_p^* \le \gamma z_i^* \le 1$. Furthermore, the solution $(\gamma z^*, \gamma\beta^*, \gamma t^*)$ has a strictly smaller objective value than the solution (z^*, β^*, t^*), which is a contradiction.

Finally, let's consider the case where $\sum_{i\in[p-1]} z_i^* - (p-2)z_p^* \ge 1$. In this case, because the constraint $(\mathbf{1}^\top\beta^*)^2 \le t$ is active, the optimal value is attained when $\mathbf{1}^\top\beta^* = -\frac{\kappa}{2}$ and $t^* = (\mathbf{1}^\top\beta^*)^2$ and (11) has the same optimal value as

$$\begin{aligned} \min \quad & a^\top z - \frac{\kappa^2}{4} \\ \text{subject to} \quad & \sum_{i\in[p-1]} z_i - (p-2)z_p \ge 1 \\ & z_p \le z_i, \quad \forall i \in [p-1] \\ & z_i \in [0,1]^p. \end{aligned}$$

From Lemma 1, the extreme points of this problem are integral, which completes the proof. □

3 Computations

We provide preliminary computations of the proposed strengthening derived in §2.2 with *hierarchy* constraints [11]. Specifically, there is a set of 3-tuples H, such that if $(i,j,k) \in H$ then $y_k \neq 0$ implies $y_i \neq 0$ and $y_j \neq 0$, resulting in

the optimization problem [30]

$$\min \frac{1}{2}\|y - X\beta\|_2^2 + \lambda \sum_{i=1}^{p} z_i \tag{12a}$$

$$\text{s.t. } \beta_i(1 - z_i) = 0 \qquad \forall i \in [p] \tag{12b}$$

$$z_k \leq z_i,\ z_k \leq z_j \qquad \forall (i,j,k) \in H \tag{12c}$$

$$z \in \{0,1\}^p, \tag{12d}$$

with L0 regularization with parameter $\lambda > 0$. We consider the following strong (and big-M free) semi-definite relaxations of (12):

•**Dynamic perspective relaxation (persp)** The dynamic perspective reformulation was proposed in [19] and involves the introduction of additional variables $B = \beta\beta^\top$:

$$\min \frac{1}{2}\|y\|_2^2 - y^\top X\beta + \frac{1}{2}\langle X^\top X, B\rangle + \lambda \sum_{i=1}^{p} z_i \tag{13a}$$

$$\text{s.t. } z_k \leq z_i,\ z_k \leq z_j \qquad \forall (i,j,k) \in H \tag{13b}$$

$$\beta_i^2 \leq z_i B_{ii} \qquad \forall i \in [p] \tag{13c}$$

$$\begin{pmatrix} 1 & \beta^\top \\ \beta & B \end{pmatrix} \succeq 0 \tag{13d}$$

$$z \in [0,1]^p. \tag{13e}$$

It corresponds to the best perspective reformulation that can be attained by decomposing the matrix $X^\top X = D + R$ with $D, R \succeq 0$ and D diagonal, and using the perspective reformulation to strengthen the term $\beta^\top D\beta$. This relaxation, depending on the diagonal dominance of matrix $X^\top X$, can be substantially stronger than the natural convex relaxation of (12). In light of Theorem 1 – and since the constraints (12c) are totally unimodular –, this formulation cannot be strengthened unless non-separable quadratic terms are accounted for.

•**Two-dimensional rank-one relaxation (R1)** The rank-one relaxation, proposed in [4], is a strengthening of the perspective reformulation by optimally decomposing $X^\top X = T + R$ where T is a sum of low-dimensional rank-one matrices, and using perspective and rank-one strengthening to strengthen the term $\beta^\top T\beta$. If all rank-one matrices are two-dimensional, then the resulting formulation involves the addition of constraints

$$\begin{pmatrix} z_i + z_j & \beta_i & \beta_j \\ \beta_i & B_{ii} & B_{ij} \\ \beta_j & B_{ij} & B_{jj} \end{pmatrix} \succeq 0$$

for all $i < j$. Observe that this formulation requires adding $O(p^2)$ constraints.

•**Hierarchical strengthening (Hier)** Corresponds to strengthening the rank-one formulation by exploiting constraints (Theorem 4). Suppose that $(i, \ell, j) \in H$; then using the same techniques used in [4], we obtain three valid inequalities

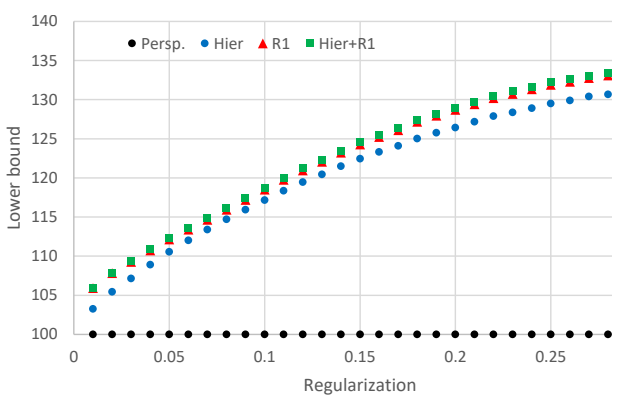
$$\begin{pmatrix} z_i & \beta_i & \beta_j \\ \beta_i & B_{ii} & B_{ij} \\ \beta_j & B_{ij} & B_{jj} \end{pmatrix} \succeq 0,\ \begin{pmatrix} z_\ell & \beta_\ell & \beta_j \\ \beta_\ell & B_{\ell\ell} & B_{\ell j} \\ \beta_j & B_{\ell j} & B_{jj} \end{pmatrix} \succeq 0 \text{ and } \begin{pmatrix} z_i + z_\ell - z_j & \beta_i & \beta_\ell & \beta_j \\ \beta_i & B_{ii} & B_{i\ell} & B_{ij} \\ \beta_\ell & B_{i\ell} & B_{\ell\ell} & B_{\ell j} \\ \beta_j & B_{ij} & B_{\ell j} & B_{jj} \end{pmatrix} \succeq 0. \tag{14}$$

The hierarchical strengthening corresponds to adding constraints (14) for every element in H to formulation (13); it requires adding only $O(p)$ constraints.

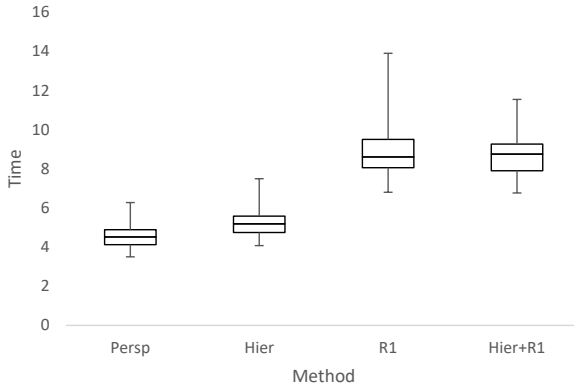
•**Full strengthening (Hier+R1)** Corresponds to adding all constraints from the rank-one relaxation and the hierarchical strengthening.

Results We compare the strength of the formulations as well as the time required to solve the SDP relaxations on the Diabetes dataset [4, 10, 21] which involves second order interactions between variables ($p = 64$). Figure 1 depicts the result. Figure 1(a) shows the optimal objective values of the different convex relaxations of (12) as a function of λ, thus larger values indicate stronger (and better) relaxations. Figure 1(b) depicts the time required to solve the relaxations; we did not observe any correlation between the value of λ and the time required, so we report aggregated times across all values of λ tested.

We observe that just using the hierarchical strengthening (`Hier`) achieves almost the same improvement in terms of the lower bound as that using the rank-one strengthening, despite only requiring $O(p)$ additional constraints instead



(a) Lower bound from the convex relaxation as a function of the regularization parameter λ.



(b) Time (in seconds).

Figure 1: Lower bounds obtained from the convex relaxation and time required to solve the relaxation. In (a) the values were scaled so that the objective value obtained from the perspective relaxation [19] is 100.

of $O(p^2)$. Indeed we observe from Figure 1(b) that the Hierarchical strengthening results in only a modest increase in the computational time with respect to the perspective relaxation, while the rank-one strengthening requires almost double that time. In addition, if the Hierarchical strengthening is used on top of the rank-one strengthening (`Hier+R1`), then we notice a small but noticeable improvement in the quality of the lower bound across all values of λ, with no apparent increase in the computational time. These preliminary results suggest that by exploiting the constraints of the optimization problems, it is possible to achieve stronger relaxations without substantially increasing the difficulty of solving the convex problems.

Appendix

Lemma 1. Suppose z^* is an extreme point of Q_g and z^* has a fractional entry. If $\sum_{i\in[p-1]} z_i^* - (p-2)z_p^* > 1$, let us consider the two cases where $z_p^* = 0$ and $z_p^* > 0$. When $z_p^* = 0$ and there exists a fractional coordinate z_i^* where $i \in [p-1]$, we can perturb z_i^* by a sufficient small quantity ϵ such that $z^* + \epsilon e_i$ and $z^* - \epsilon e_i$ are in Q_g. Then, $z^* = \frac{1}{2}(z^* + \epsilon e_i) + \frac{1}{2}(z^* - \epsilon e_i)$ which contradicts the fact that z^* is an extreme point of Q_g. When $1 > z_p^* > 0$ we can perturb z_p^* and all other z_i^*'s with $z_i^* = z_p^*$ by a sufficiently small quantity ϵ and stay in Q_g. Similarly, we will reach a contradiction.

Now suppose $\sum_{i\in[p-1]} z_i^* - (p-2)z_p^* = 1$, and let us consider again the two cases where $z_p^* = 0$ and $z_p^* > 0$. When $z_p^* = 0$, $z^* = z_1^* e_1 + \cdots + z_{(p-1)}^* e_{(p-1)}$, which is a contradiction since we can write z^* as a convex combination of points $e_i \in Q_g, i \in [p-1]$ and there exists at least two indices $i, j \in [p-1], i \neq j$ such that $1 > z_i^*, z_j^* > 0$ by the fact that z^* has a fractional entry and $\sum_{i\in[p-1]} z_i^* = 1, 0 \leq z_i^* \leq 1, \forall i$. When $1 > z_p^* > 0$, we first show that there exists at most one 1 in $z_1^*, z_2^*, \ldots, z_{(p-1)}^*$. Suppose we have $z_i^* = 1$ and $z_j^* = 1$ for $i, j \in [p-1]$ with $i \neq j$, then $\sum_{i\in[p-1]} z_i^* - (p-2)z_p^* = z_i^* + \sum_{l\in[p-1], l\neq i}(z_l^* - z_p^*) \geq z_i^* + (z_j^* - z_p^*) > z_i^* = 1$, which is a contradiction. We now show that we can perturb z_p^* and the $p-2$ smallest elements in $z_i^*, i \in [p-1]$ by a small quantity ϵ and remain in Q_g. The equality $\sum_{i\in[p-1]} z_i - (p-2)z_p = 1$ clearly holds after the perturbation. And, adding a small quantity ϵ to z_p^* and the $p-2$ smallest elements in $z_i^*, i \in [p-1]$ will not violate the hierarchy constraint since the largest element in $z_i^*, i \in [p-1]$ has to be strictly greater than z_p^*. (Note that if $z_i^* = z_p^*, \forall i \in [p]$, $\sum_{i\in[p-1]} z_i^* - (p-2)z_p^* = z_p^* < 1$.) Since $z_i^* \geq z_p^* > 0, \forall i \in [p-1]$ subtracting a small quantity ϵ will not violate the non-negativity constraint. Thus, we can write z^* as a convex combination of two points in Q_g, which is a contradiction. □

References

[1] Aktürk, M. S., Atamtürk, A., and Gürel, S. (2009). A strong conic quadratic reformulation for machine-job assignment with controllable processing times. *Operations Research Letters*, 37(3):187–191.

[2] Anstreicher, K. M. (2012). On convex relaxations for quadratically constrained quadratic programming. *Mathematical Programming*, 136(2):233–251.

[3] Atamtürk, A. and Gómez, A. (2018). Strong formulations for quadratic optimization with M-matrices and indicator variables. *Mathematical Programming*, 170(1):141–176.

[4] Atamtürk, A. and Gómez, A. (2019). Rank-one convexification for sparse regression. *http://www.optimization-online.org/DB_HTML/2019/01/7050.html*.

[5] Atamtürk, A., Gómez, A., and Han, S. (2018). Sparse and smooth signal estimation: Convexification of L0 formulations. *http://www.optimization-online.org/DB_HTML/2018/11/6948.html*.

[6] Bacci, T., Frangioni, A., Gentile, C., and Tavlaridis-Gyparakis, K. (2019). New minlp formulations for the unit commitment problems with ramping constraints. *Optimization Online*.

[7] Belotti, P., Góez, J. C., Pólik, I., Ralphs, T. K., and Terlaky, T. (2015). A conic representation of the convex hull of disjunctive sets and conic cuts for integer second order cone optimization. In *Numerical Analysis and Optimization*, pages 1–35. Springer.

[8] Bertsimas, D., Cory-Wright, R., and Pauphilet, J. (2019). A unified approach to mixed-integer optimization: Nonlinear formulations and scalable algorithms. *arXiv preprint arXiv:1907.02109*.

[9] Bertsimas, D. and King, A. (2016). OR Forum – An algorithmic approach to linear regression. *Operations Research*, 64(1):2–16.

[10] Bertsimas, D., King, A., and Mazumder, R. (2016). Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852.

[11] Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of Statistics*, 41(3):1111.

[12] Bienstock, D. and Michalka, A. (2014). Cutting-planes for optimization of convex functions over nonconvex sets. *SIAM Journal on Optimization*, 24(2):643–677.

[13] Burer, S. and Kılınç-Karzan, F. (2017). How to convexify the intersection of a second order cone and a nonconvex quadratic. *Mathematical Programming*, 162(1-2):393–429.

[14] Ceria, S. and Soares, J. (1999). Convex programming for disjunctive convex optimization. *Mathematical Programming*, 86:595–614.

[15] Cozad, A., Sahinidis, N. V., and Miller, D. C. (2014). Learning surrogate models for simulation-based optimization. *AIChE Journal*, 60(6):2211–2227.

[16] Cozad, A., Sahinidis, N. V., and Miller, D. C. (2015). A combined first-principles and data-driven approach to model building. *Computers & Chemical Engineering*, 73:116–127.

[17] Dong, H. (2019). On integer and MPCC representability of affine sparsity. *Operations Research Letters*, 47(3):208–212.

[18] Dong, H., Ahn, M., and Pang, J.-S. (2019). Structural properties of affine sparsity constraints. *Mathematical Programming*, 176(1-2):95–135.

[19] Dong, H., Chen, K., and Linderoth, J. (2015). Regularization vs. relaxation: A conic optimization perspective of statistical variable selection. *arXiv preprint arXiv:1510.06083*.

[20] Dong, H. and Linderoth, J. (2013). On valid inequalities for quadratic programming with continuous variables and binary indicators. In *International Conference on Integer Programming and Combinatorial Optimization*, pages 169–180. Springer.

[21] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

[22] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

[23] Frangioni, A., Furini, F., and Gentile, C. (2016). Approximated perspective relaxations: a project and lift approach. *Computational Optimization and Applications*, 63(3):705–735.

[24] Frangioni, A. and Gentile, C. (2006). Perspective cuts for a class of convex 0–1 mixed integer programs. *Mathematical Programming*, 106:225–236.

[25] Frangioni, A. and Gentile, C. (2007). SDP diagonalizations and perspective cuts for a class of nonseparable MIQP. *Operations Research Letters*, 35(2):181–185.

[26] Frangioni, A., Gentile, C., Grande, E., and Pacifici, A. (2011). Projected perspective reformulations with applications in design problems. *Operations Research*, 59(5):1225–1232.

[27] Frangioni, A., Gentile, C., and Hungerford, J. (2019). Decompositions of semidefinite matrices and the perspective reformulation of nonseparable quadratic programs. *Mathematics of Operations Research*.

[28] Günlük, O. and Linderoth, J. (2010). Perspective reformulations of mixed integer nonlinear programs with indicator variables. *Mathematical Programming*, 124:183–205.

[29] Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. CRC press.

[30] Hazimeh, H. and Mazumder, R. (2019). Learning hierarchical interactions at scale: A convex optimization approach. *arXiv preprint arXiv:1902.01542*.

[31] Hijazi, H., Bonami, P., Cornuéjols, G., and Ouorou, A. (2012). Mixed-integer nonlinear programs featuring "on/off" constraints. *Computational Optimization and Applications*, 52(2):537–558.

[32] Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical science: A Review Journal of the Institute of Mathematical Statistics*, 27(4).

[33] Jeon, H., Linderoth, J., and Miller, A. (2017). Quadratic cone cutting surfaces for quadratic programs with on–off constraints. *Discrete Optimization*, 24:32–50.

[34] Kılınç-Karzan, F. and Yıldız, S. (2014). Two-term disjunctions on the second-order cone. In *International Conference on Integer Programming and Combinatorial Optimization*, pages 345–356. Springer.

[35] Manzour, H., Küçükyavuz, S., and Shojaie, A. (2019). Integer programming for learning directed acyclic graphs from continuous data. *arXiv preprint arXiv:1904.10574*.

[36] Miller, A. (2002). *Subset selection in regression*. Chapman and Hall/CRC.

[37] Modaresi, S., Kılınç, M. R., and Vielma, J. P. (2016). Intersection cuts for nonlinear integer programming: Convexification techniques for structured sets. *Mathematical Programming*, 155(1-2):575–611.

[38] Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234.

[39] Richard, J.-P. P. and Tawarmalani, M. (2010). Lifting inequalities: a framework for generating strong cuts for nonlinear programs. *Mathematical Programming*, 121(1):61–104.

[40] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, pages 267–288.

[41] Vielma, J. P. (2019). Small and strong formulations for unions of convex sets from the Cayley embedding. *Mathematical Programming*, 177(1-2):21–53.

[42] Wang, A. L. and Kilinc-Karzan, F. (2019). The generalized trust region subproblem: solution complexity and convex hull results. *arXiv preprint arXiv:1907.08843*.

[43] Wang, A. L. and Kılınç-Karzan, F. (2019). On the tightness of sdp relaxations of qcqps. *Optimization Online*.

[44] Wu, B., Sun, X., Li, D., and Zheng, X. (2017). Quadratic convex reformulations for semicontinuous quadratic programming. *SIAM Journal on Optimization*, 27(3):1531–1553.

[45] Xie, W. and Deng, X. (2018). The CCP selector: Scalable algorithms for sparse ridge regression from chance-constrained programming. *arXiv preprint arXiv:1806.03756*.

[46] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38:894–942.

[47] Zheng, X., Sun, X., and Li, D. (2014). Improving the performance of MIQP solvers for quadratic programs with cardinality and minimum threshold constraints: A semidefinite program approach. *INFORMS Journal on Computing*, 26(4):690–703.