COMPACT EXTENDED FORMULATIONS FOR LOW-RANK FUNCTIONS WITH INDICATOR VARIABLES

SHAONING HAN AND ANDRÉS GÓMEZ

October 2021

Abstract. We study the mixed-integer epigraph of a low-rank convex function with non-convex indicator constraints, which are often used to impose logical constraints on the support of the solutions. Extended formulations describing the convex hull of such sets can easily be constructed via disjunctive programming, although a direct application of this method often yields prohibitively large formulations, whose size is exponential in the number of variables. In this paper, we propose a new disjunctive representation of the sets under study, which leads to compact formulations with size exponential in the rank of the function, but polynomial in the number of variables. Moreover, we show how to project out the additional variables for the case of rank-one functions, recovering or generalizing known results for the convex hulls of such sets (in the original space of variables).

Keywords. Mixed-integer nonlinear optimization, convexification, disjunctive programming, indicator variables.

1. Introduction

In this paper, we consider a general mixed-integer convex optimization problem with indicator variables

$$\min_{x,z} \left\{ F(x) : (x,z) \in \mathcal{F} \subseteq \mathbb{R}^n \times \{0,1\}^n, x_i(1-z_i) = 0 \; \forall i \in [n] \right\}, \quad (1)$$

where $F : \mathbb{R}^n \to \mathbb{R}$ is a convex function, $\mathcal{F}$ is the feasible region, and $[n] \stackrel{\text{def}}{=} \{1, 2, \ldots, n\}$. Each binary variable z_i indicates whether a given continuous variable x_i is zero or not. In other words, $z_i = 0 \implies x_i = 0$, and $z_i = 1$ allows x_i to take any value. Optimization problem (1) arises in a variety of settings, including best subset selection problems in statistics [14], and portfolio optimization problems in finance [15].

S. Han, A. Gómez: Daniel J. Epstein Department of Industrial and Systems Engineering, Viterbi School of Engineering, University of Southern California, CA 90089. shaoning@usc.edu, gomezand@usc.edu.

In practice, the objective function often has the form $F(x) = \sum_{i\in[m]} f_i(x)$, where each $f_i(x)$ is a composition of a relatively simple convex function and a low-rank linear map. This observation motivates the need for a comprehensive study of the mixed-integer set

$$\mathcal{Q} \stackrel{\text{def}}{=} \left\{ (t,x,z) \in \mathbb{R}^{n+1} \times \{0,1\}^n : \begin{array}{l} t \geq f(x), x_i \geq 0 \; \forall i \in \mathcal{I}_+, \\ z \in \{0,1\}^n, x_i(1-z_i) = 0 \; \forall i \in [n] \end{array} \right\}, \tag{2}$$

where $f(x) = g(Ax) + c^\top x$, $g : \mathbb{R}^k \to \mathbb{R}$ is a proper closed convex function, A is a $k \times n$ matrix, $c \in \mathbb{R}^n$, and $\mathcal{I}_+ \subseteq [n]$ is the subset of variables restricted to be non-negative.

Disjunctive programming is a powerful modeling tool to represent a non-convex optimization problem in the form of disjunctions, especially when binary variables are introduced to encode the logical conditions such as sudden changes, either/or decisions, implications, etc. [10]. The theory of linear disjunctive programming is first pioneered by Egon Balas in 1970s [9, 7, 8, 11], and later extended to the nonlinear case [17, 37, 33, 25, 12, 16, 31, 32, 34, 18, 40]. Once a mixed-integer set is modeled as a collection of disjunctive sets, its convex hull can be described easily as a Minkowski sum of the scaled convex sets with each obtained by creating a copy of original variables. Such extended formulations are the strongest possible for the mixed-integer set under study. However, a potential downside of such formulations is that, often, the number of additional variables required in the description of the convex hull is exponential in the number of binary variables. Thus, a direct application of disjunctive programming as mentioned can be unfavorable in practice.

A possible approach to implement the extended formulation induced by disjunctive programming in an efficient way is to reduce the number of additional variables introduced in the model without diminishing the relaxation strength. In principle, this goal can be achieved by means of Fourier-Motzkin elimination [19]. This method is practical if the set under study can be naturally expressed using few disjunctions, e.g., to describe piecewise linear functions [2, 28], involving few binary variables [4, 23], or separable [27]. However, projecting out variables in a moderate-size model can be very challenging [3], not to mention in a high-dimensional setting.

Regarding the nonlinear set with indicator variables $\mathcal{Q}$, in the simplest case where the MINLP is separable, it is known that the convex hull can be described by the perspective function of the objective function supplemented by the constraints defining the feasible region [1, 20, 21, 22, 27, 29, 39, 13].

Two papers [38] and [6] are closely related to our work. The mixed-integer sets studied in both can be viewed as a special realization of $\mathcal{Q}$ in (2). Wei

et al. [38] characterize the closure of the convex hull if $\mathcal{Q}$ when the defining function g is a univariate function and the continuous variables are free, i.e. $\mathcal{I}_+ = \emptyset$. Atamtürk and Gómez [6] study the setting with sign-constrained continuous variables, and univariate quadratic functions g: in such cases, the sign-restrictions resulted in more involved structures for the closure of the convex hull of $\mathcal{Q}$.

Contributions and outline. In this paper we show how to construct extended formulations of set $\mathcal{Q}$, requiring at most $\mathcal{O}(n^k)$ copies of the variables. In particular, if the rank k is small, then the resulting formulations are indeed much smaller than those resulting from a natural application of disjunctive programming. Moreover, for the special case of $k = 1$, we show how we are able to recover and improve existing results in the literature, either by providing smaller extended formulations (linear in n), or by providing convexifications in the original space of variables for a more general class of functions.

The rest of the paper is organized as follows. In Section 2, we provide relevant background and the main result of the paper: a compact extended formulation of $\mathcal{Q}$. In Section 3, using the results in Section 2, we derive the explicit form of the convex hull of the rank-one set $\mathcal{Q}$ in the original state of variables. In Section 4, we present the complexity results showing that tractable convexifications of $\mathcal{Q}$ are unlikely if additional constraints are imposed on the continuous variables. Finally, in Section 5, we conclude the paper.

2. A convex analysis perspective on convexification

In this section, we first introduce necessary preliminaries in convex analysis and notations adopted in this paper. After that we present our main results and their connections with previous works in literature.

2.1. Notations and preliminaries.

Throughout this paper, we assume $f(x)$ a proper closed convex function from $\mathbb{R}^n$ to $\mathbb{R} \cup \{+\infty\}$. If $f(x) < +\infty$ for all x, $f(x)$ is called *finite*. We denote the *effective domain* of f by $\mathrm{dom}(f)$ and the convex *conjugate* of $f(\cdot)$ by $f^*(\cdot)$ which is defined as

$$f^*(\alpha) \stackrel{\text{def}}{=} \max_x \ \alpha^\top x - f(x).$$

The *perspective function* of $f(\cdot)$ is defined as

$$f^\pi(x, \lambda) \stackrel{\text{def}}{=} \begin{cases} \lambda f\left(\frac{x}{\lambda}\right) & \text{if } \lambda \geq 0 \\ +\infty & \text{o.w.,} \end{cases}$$

where $0f(x/0)$ is interpreted as the *recession function* of f at x. By this definition, function $f^\pi(\cdot)$ is homogeneous, closed and convex; see Section 8

in [35]. Moreover, we borrow the term *rank* which is normally defined for an affine mapping and extend its definition to a general nonlinear convex function.

Definition 1 (Rank of convex functions)**.** Given a proper closed convex function $f(x)$, the rank[1] $\operatorname{rank}(f)$ is defined as the smallest integer k such that $f(x)$ can be expressed in the form $f(x) = g(Ax) + c^\top x$ for some closed convex function $g(\cdot)$, a $k \times n$ matrix A and a vector $c \in \mathbb{R}^n$.

For example, the rank of an affine function $f(x) = c^\top x$ is simply 0. The rank of a convex quadratic function $f(x) = x^\top Ax$ coincides with $\operatorname{rank}(A)$, where $A \succeq 0$ is a positive semidefinite (PSD) matrix.

We let $\mathbf{e}$ be the vector of all ones (whose dimension can be inferred from the context). For a set $\mathcal{S}$, we denote the convex hull of S as $\operatorname{conv}(\mathcal{S})$, and its closure as $\operatorname{cl conv}(\mathcal{S})$. For any scalar $\lambda_1 \geq 0$ and two generic sets $\mathcal{S}_1$, $\mathcal{S}_2$ in a proper Euclidean space, we define $\lambda\mathcal{S}_1 \stackrel{\text{def}}{=} \{\lambda x : x \in \mathcal{S}_1\}$ and $\mathcal{S}_1 + \mathcal{S}_2 \stackrel{\text{def}}{=} \{x^1 + x^2 : x^1 \in \mathcal{S}_1, x^2 \in \mathcal{S}_2\}$ is the Minkowski sum. We denote the *indicator function* of $\mathcal{S}$ by $\delta(x; \mathcal{S})$, which is defined as $\delta(x; \mathcal{S}) = 0$ if $x \in \mathcal{S}$ and $\delta(x; \mathcal{S}) = +\infty$ otherwise. By the above notations, the convex conjugate of $\delta(x; \mathcal{S})$ is

$$\delta^*(\alpha; \mathcal{S}) = \max_x \ \alpha^\top x - \delta(x; \mathcal{S}) = \max_{x \in \mathcal{S}} \ \alpha^\top x,$$

which is known as the *support function* of $\mathcal{S}$. To be consistent with the definition of $0f(x/0)$, we interpret $0\mathcal{S}$ as the *recession cone* of $\mathcal{S}$ throughout this paper. The derivation of the succeeding work relies on the following one-to-one correspondence between closed convex sets and support functions, whose proof can be found in classical books of convex analysis, e.g. Section 13 in [35] and Chapter C.2 in [30].

Proposition 1. *Given a set $\mathcal{S}$ and a closed convex set $\mathcal{T}$, $\mathcal{T} = \operatorname{cl conv}(\mathcal{S})$ if and only if $\delta^*(\cdot; \mathcal{T}) = \delta^*(\cdot; \mathcal{S})$.*

For convenience, we repeat the set of interest:

$$\mathcal{Q} \stackrel{\text{def}}{=} \left\{ (t, x, z) \in \mathbb{R}^{n+1} \times \{0,1\}^n : \begin{array}{l} t \geq f(x), x_i \geq 0 \ \forall i \in \mathcal{I}_+, \\ x_i(1 - z_i) = 0 \ \forall i \in [n] \end{array} \right\},$$

where $f(x) = g(Ax) + c^\top x$ is a proper closed convex function.

Note that if the complementary constraints $x_i(1 - z_i) = 0$ are removed from $\mathcal{Q}$, then x and z are decoupled and $\operatorname{cl conv}(\mathcal{Q})$ reduces simply to $\mathcal{X} \times [0,1]^n$, where

$$\mathcal{X} \stackrel{\text{def}}{=} \left\{ (t, x) \in \mathbb{R}^{n+1} : t \geq f(x), x_i \geq 0 \ \forall i \in \mathcal{I}_+ \right\}.$$

[1]The definition of $\operatorname{rank}(f)$ is different from the one adopted in classical convex analysis (see Section 8 in [35]). If $\operatorname{dom}(f)$ is full-dimensional, the two definitions coincide.

For the purpose of decomposing cl conv($\mathcal{Q}$), for any $\mathcal{I} \subseteq [n]$, define the following sets:

$$\begin{aligned}
\mathcal{X}(\mathcal{I}) &\stackrel{\text{def}}{=} \mathcal{X} \cap \{(t,x) : x_i = 0 \; \forall i \notin \mathcal{I}\},\\
\mathcal{Z}(\mathcal{I}) &\stackrel{\text{def}}{=} \{z \in \{0,1\}^n : z_i = 1 \; \forall i \in \mathcal{I}\},\\
\mathcal{V}(\mathcal{I}) &\stackrel{\text{def}}{=} \mathcal{X}(\mathcal{I}) \times \mathcal{Z}(\mathcal{I}).
\end{aligned}$$

Notice that for any $(t,x,z) \in \mathcal{V}(\mathcal{I})$, either $x_i = 0$ or $z_i = 1$ $\forall i \in [n]$. Therefore, $x_i(1-z_i) = 0 \; \forall i \in [n]$ and thus $\mathcal{V}(\mathcal{I}) \subseteq \mathcal{Q}$. Furthermore, $\mathcal{Q}$ can be expressed as a disjunction

$$\mathcal{Q} = \bigcup_{\mathcal{I} \subseteq [n]} \mathcal{V}(\mathcal{I}). \tag{3}$$

Note that usual disjunctive programming techniques are based on (3), creating copies of variables for every $I \subseteq [n]$, resulting in an exponential number of variables.

Finally, define

$$\mathcal{R} \stackrel{\text{def}}{=} \{(t,x,z) : t \geq 0,\; Ax = 0, x_i \geq 0,\; \forall i \in \mathcal{I}_+,\; z_i = 1,\; \forall i \in [n]\},$$

which is a closed convex set. For any $y \in \mathbb{R}^n$, we denote the *support* of y by $\text{supp}(y) \stackrel{\text{def}}{=} \{i : y_i \neq 0\}$

2.2. **Convex hull characterization.** In this section, we characterize cl conv(Q). We first show that if $f(x)$ is homogeneous, under mild conditions, cl conv($\mathcal{Q}$) is simply the natural relaxation of $\mathcal{Q}$.

Proposition 2. *If $f(x)$ is a homogeneous function, then cl conv($\mathcal{Q}$) = $\mathcal{X} \times [0,1]^n$.*

Proof. Since f is homogeneous, each set $\mathcal{X}(\mathcal{I})$ is a closed convex cone. Moreover, $\mathcal{I}_1 \subseteq \mathcal{I}_2$ implies that $\mathcal{X}(\mathcal{I}_1) \subseteq \mathcal{X}(\mathcal{I}_2)$, which further implies that

$$\mathcal{X}(\mathcal{I}_2) \subseteq \mathcal{X}(\mathcal{I}_1) + \mathcal{X}(\mathcal{I}_2) \subseteq \mathcal{X}(\mathcal{I}_2) + \mathcal{X}(\mathcal{I}_2) = \mathcal{X}(\mathcal{I}_2), \tag{4}$$

where the last equality results from that $\mathcal{X}(\mathcal{I}_2)$ is conic. Hence, equality holds throughout (4). By (3),

$$\begin{aligned}
&\text{cl conv}(\mathcal{Q}) = \text{cl conv}\left(\bigcup_{\mathcal{I}\subseteq[n]} \mathcal{V}(\mathcal{I})\right)\\
=&\text{cl conv}\left(\bigcup_{\mathcal{I}\subseteq[n]} \mathcal{X}(\mathcal{I}) \times \text{conv}(\mathcal{Z}(\mathcal{I}))\right) \quad (\mathcal{X}(\mathcal{I}) \text{ is convex; } \mathcal{X}, \mathcal{Z} \text{ are decoupled})\\
=&\bigcup_{\lambda\in\mathbb{R}^{2^n}_+ : \mathbf{e}^\top\lambda=1} \sum_{\mathcal{I}\subseteq[n]} \lambda_{\mathcal{I}} \left(\mathcal{X}(\mathcal{I}) \times \text{conv}(\mathcal{Z}(\mathcal{I}))\right)\\
=&\bigcup_{\lambda\in\mathbb{R}^{2^n}_+ : \mathbf{e}^\top\lambda=1} \mathcal{X}([n]) \times \sum_{\mathcal{I}\subseteq[n]} \lambda_{\mathcal{I}}\text{conv}(\mathcal{Z}(\mathcal{I})) \qquad (\mathcal{X}(\mathcal{I}) \text{ is conic and (4)})\\
=&\mathcal{X} \times \bigcup_{\lambda\in\mathbb{R}^{2^n}_+ : \mathbf{e}^\top\lambda=1} \sum_{\mathcal{I}\subseteq[n]} \lambda_{\mathcal{I}}\text{conv}(\mathcal{Z}(\mathcal{I})) \qquad (\mathcal{X}=\mathcal{X}([n]))\\
=&\mathcal{X} \times [0,1]^n.
\end{aligned}$$

□

Proposition 2 generalizes Proposition 1 of [24] (where f is the ℓ_2 norm). Next, we present the main result of the paper, characterizing cl conv($\mathcal{Q}$) without the assumption of homogeneity. In particular, we show that cl conv($\mathcal{Q}$) can be constructed from substantially less disjunctions than those given in (3).

Theorem 1. *Assume* $\text{rank}(f) \leq k$ *and* $f(0) = 0$. *Then*

$$\textit{cl conv}(\mathcal{Q}) = \textit{cl conv}\left(\left(\bigcup_{\mathcal{I}:|\mathcal{I}|\leq k} \mathcal{V}(I)\right) \cup \mathcal{R}\right).$$

Moreover, if $\{x \in \mathbb{R}^n : Ax = 0,\ x_i \geq 0\ \forall i \in \mathcal{I}_+\} = \{0\}$, *then* $\mathcal{R}$ *can be removed from the disjunction.*

Informally, $\bigcup_{\mathcal{I}:|\mathcal{I}|\leq k} \mathcal{V}(I)$ and $\mathcal{R}$ correspond to the "extreme points" and "extreme rays" of cl conv($\mathcal{Q}$), respectively. From (3), $\mathcal{Q}$ is a disjunction of exponentially many pieces of $\mathcal{V}(\mathcal{I})$. However, given a low-rank function f, Theorem 1 states that cl conv($\mathcal{Q}$) can be generated (using disjunctive programming) from a much smaller number of sets $\mathcal{V}(\mathcal{I})$. We also remark that condition $f(0) = 0$ plays a minor role in the derivation of Theorem 1. If $0 \in \text{dom}(f)$ but $f(0) \neq 0$, one can study $f(x) - f(0)$.

Proof of Theorem 1. Denote $\mathcal{S} = \bigcup_{\mathcal{I}:|\mathcal{I}|\leq k} \mathcal{V}(I) \cup \mathcal{R}$. Since $\mathcal{V}(\mathcal{I}) \subseteq \mathcal{Q}$ for all $\mathcal{I}$, and $\mathcal{R} \subseteq \mathcal{Q}$, we find that $\mathcal{S} \subseteq \mathcal{Q}$. Thus, $\delta^*(\cdot; \text{cl conv}(\mathcal{S})) = \delta^*(\cdot; \mathcal{S}) \leq \delta^*(\cdot; \mathcal{Q})$.

Due to Proposition 1, it remains to prove the opposite direction, namely that $\delta^*(\cdot;\mathcal{S}) \geq \delta^*(\cdot;\mathcal{Q})$.

Since $\operatorname{rank}(f) \leq k$, there exists $g(\cdot)$, $A \in \mathbb{R}^{k\times n}$ and c such that $f(x) = g(Ax) + c^\top x$. Taking any $\alpha = (\alpha_t, \alpha_x, \alpha_z) \in \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n$, if $\alpha_t > 0$, $\delta^*(\cdot;\mathcal{S}) = \delta^*(\cdot;\mathcal{Q}) = +\infty$ because t is unbounded from above. We now assume $\alpha_t \leq 0$, and define

$$\ell(x,z) \stackrel{\text{def}}{=} \alpha_t g(Ax) + (\alpha_x + \alpha_t c)^\top x + \alpha_z^\top z.$$

Then $\delta^*(\alpha;\mathcal{S}) = \max\{\ell(x,z) : (t,x,z) \in \mathcal{S}\}$ and $\delta^*(\alpha;\mathcal{Q}) = \max\{\ell(x,z) : (t,x,z) \in \mathcal{Q}\}$. Given any fixed $(\bar{t},\bar{x},\bar{z}) \in \mathcal{Q}$, consider the linear program

$$\begin{aligned} v^* \stackrel{\text{def}}{=} \max_{x\in\mathbb{R}^n}\ & (\alpha_x + \alpha_t c)^\top x \\ \text{s.t.}\ & Ax = A\bar{x} \\ & \bar{x}_i x_i \geq 0 \qquad \forall i \in \operatorname{supp}(\bar{x}) \subseteq \operatorname{supp}(\bar{z}) \\ & x_i = 0 \qquad\quad \forall i \in [n]\backslash \operatorname{supp}(\bar{x}). \end{aligned} \tag{5}$$

Note that linear program (5) is always feasible setting $x = \bar{x}$. Moreover, every feasible solution (or direction) x of (5) satisfies $x_i(1-\bar{z}_i) = 0$.

If $v^* = +\infty$, then there exists a feasible direction d such that $Ad = 0$ and $(\alpha_x + \alpha_t c)^\top d > 0$. It implies that for any $r \geq 0$, $f(rd) = g(rAd) + rc^\top d = g(0) + rc^\top d = rc^\top d$. Hence, $(rc^\top d, rd, \mathbf{e}) \in \mathcal{R} \subseteq Q$. Furthermore,

$$\ell(rd,\mathbf{e}) = (\alpha_x + \alpha_t c)^\top dr + \alpha_z^\top \mathbf{e} \to +\infty, \qquad (\text{as } r \to +\infty)$$

which implies $\delta^*(\alpha;\mathcal{S}) = \delta^*(\alpha;\mathcal{Q}) = +\infty$.

If v^* is finite, there exists an optimal solution x^* to (5). Moreover, x^* can be taken as an extreme point of the feasible region of (5) which is a pointed polytope. It implies that n linearly independent constraints must be active at x^*. Since $\operatorname{rank}(A) \leq k$, at least $n-k$ constraints of the form $\bar{x}_i x_i \geq 0$ hold at equality. Namely, x^* satisfies $|\operatorname{supp}(x^*)| \leq k$. Since $A\bar{x} = Ax^*$, we can define

$$\begin{aligned} t^* \stackrel{\text{def}}{=}& \bar{t} + c^\top x^* - \bar{x} \\ \geq& f(\bar{x}) + c^\top x^* - \bar{x} \\ =& g(A\bar{x}) + c^\top \bar{x} + c^\top x^* - \bar{x} \\ =& g(Ax^*) + c^\top x^* = f(x^*) \end{aligned}$$

Setting $\mathcal{I} = \operatorname{supp}(x^*)$, one can deduce that $(t^*,x^*) \in \mathcal{X}(\mathcal{I})$, and $\bar{z} \in \mathcal{Z}(\mathcal{I})$ because $\operatorname{supp}(x^*) \subseteq \operatorname{supp}(\bar{z})$. Thus, $(t^*,x^*,\bar{z}) \in \mathcal{V}(\mathcal{I}) \subseteq \mathcal{S} \subseteq \mathcal{Q}$, and $(\alpha_x + \alpha_t c)^\top x^* \geq (\alpha_x + \alpha_t c)^\top \bar{x}$ indicates that $\ell(x^*,\bar{z}) \geq \ell(\bar{x},\bar{z})$. Namely, for an arbitrary point $(\bar{t},\bar{x},\bar{z}) \in \mathcal{Q}$, there always exists a point in $\mathcal{S}$ with a superior objective value of $\ell(\cdot)$. Therefore, $\delta^*(\alpha;\mathcal{S}) \geq \delta^*(\alpha;\mathcal{Q})$, completing the proof of the main conclusion.

The last statement of the theorem follows since if $\{x \in \mathbb{R}^n : Ax = 0,\ x_i \geq 0\ \forall i \in \mathcal{I}_+\} = \{0\}$, then the feasible region of (5) is bounded and thus v^* is always finite. □

Using the disjunctive representation of Theorem 1 and usual disjunctive programming techniques [17], one can immediately obtain extended formulations requiring at most $\mathcal{O}(n^k)$ copies of the variables. Moreover, it is often easy to project out some of the additional variables, resulting in formulations with significantly less variables or, in some cases, formulations in the original space of variables. We illustrate these concepts in the next section with $k = 1$.

3. Rank-one convexification

In this section, we show how to use Theorem 1 and disjunctive programming to derive convexifications for rank-one functions. In particular, throughout this section, we make the following assumptions:

Assumption 1. Function f is given by $f = g\left(\sum_{i=1}^n a_i x_i\right) + c^\top x$, where g is a finite one-dimensional function, $f(0) = g(0) = 0$, and $a_i \neq 0\ \forall i \in n$. For simplicity, we also assume $c = 0$.

First in Section 3.1 we derive an extended formulation with a polynomial number of additional variables. Then in Section 3.2 we project out the additional variables for the case with free continuous variables, and recover the results of [38]. Similarly, in Section 3.3, we provide the description of the convex hull of cases with non-negative continuous variables in the original space of variables, generalizing the results of [6] and [4] to general (not necessarily quadratic) functions g. We also show that the extended formulation proposed in this paper is more amenable to implementation that the inequalities proposed in [6].

3.1. Extended formulation of cl conv($\mathcal{Q}$).

We first discuss the compact extended formulation of cl conv($\mathcal{Q}$) that can be obtained directly for the disjunctive representation given in Theorem 1, using $2n$ additional variables.

Proposition 3. *Under Assumption 1, $(t,x,z) \in$ cl conv$(\mathcal{Q})$ if and only if there exists $\lambda, \tau \in \mathbb{R}^n$ such that the inequality system*

$$\begin{aligned} & t \geq \sum_{i=1}^{n} g^{\pi}(a_i(x_i - \tau_i), \lambda_i), \\ & a^{\top}\tau = 0,\ 0 \leq \tau_i \leq x_i\ \forall i \in \mathcal{I}_+, \\ & \lambda_i \leq z_i \leq 1\ \forall i \in [n], \\ & \lambda \geq 0,\ \sum_{i=1}^{n} \lambda_i \leq 1 \end{aligned} \tag{6}$$

is satisfied.

Proof. By Theorem 1,

$$\begin{aligned} & \text{cl conv}(\mathcal{Q}) = \text{cl conv}\left(\mathcal{V}(\emptyset) \bigcup_{i\in[n]} \mathcal{V}(\{i\}) \cup \mathcal{R}\right) \\ = & \bigcup_{\lambda \in \mathbb{R}^{n+2}_+ : \mathbf{e}^{\top}\lambda = 1} \left(\lambda_0 \text{cl conv}(\mathcal{V}(\emptyset)) + \sum_{i=1}^{n} \lambda_i \text{cl conv}(V(\{i\})) + \lambda_{[n]}\mathcal{R}\right), \end{aligned}$$

where

$$\begin{aligned} & \lambda_0 \text{cl conv}(\mathcal{V}(\emptyset)) = \{(t_0, x^0, z^0) : t_0 \geq 0,\ x^0 = 0,\ 0 \leq z^0 \leq \lambda_0\}, \\ & \lambda_i \text{cl conv}(\mathcal{V}(\{i\})) = \left\{(t_i, x^i, z^i) : \begin{array}{l} t_i \geq g^{\pi}(a_i x_i^i, \lambda_i),\ z_i^i = \lambda_i,\ 0 \leq z_j^i \leq \lambda_i\ \forall j \neq i, \\ x_i^i \geq 0 \text{ if } i \in \mathcal{I}_+,\ x_j^i = 0\ \forall j \neq i \end{array}\right\}, \\ & \lambda_{[n]}\mathcal{R} = \left\{\left(t_{[n]}, \tau, z^{[n]}\right) : t_{[n]} \geq 0,\ a^{\top}\tau = 0,\ \tau_i \geq 0\ \forall i \in \mathcal{I}_+,\ z^{[n]} = \lambda_{[n]}\mathbf{e}\right\}. \end{aligned}$$

It follows that $(t, x, z) \in \operatorname{cl}\operatorname{conv}(\mathcal{Q})$ if and only if the following inequality system has a solution:

$$
\begin{aligned}
& t \geq \sum_{i=1}^{n} g^{\pi}(a_i x_i^i, \lambda_i),\ a^{\top}\tau = 0, && \\
& x_i = x_i^i + \tau_i && \forall i \in [n], \qquad (7a)\\
& x_i^i \geq 0,\ \tau_i \geq 0 && \forall i \in \mathcal{I}_+, \\
& z_j = \sum_{i=1}^{n} z_j^i + \lambda_{[n]} + z_j^0 && \forall j \in [n], \\
& z_i^i = \lambda_i,\ 0 \leq z_i^0 \leq \lambda_0 && \forall i \in [n], \qquad (7b)\\
& 0 \leq z_j^i \leq \lambda_i && \forall j \neq i \in [n], \\
& \lambda \geq 0,\ \lambda_0 + \sum_{i=1}^{n} \lambda_i + \lambda_{[n]} = 1. &&
\end{aligned}
$$

We now show how to simply the above inequality system step by step. First, we can substitute out x_i^i and z_i^i using (7a) and (7b), obtaining the system

$$
\begin{aligned}
& t \geq \sum_{i=1}^{n} g^{\pi}(a_i(x_i - \tau_i), \lambda_i),\ a^{\top}\tau = 0, && \\
& 0 \leq \tau_i \leq x_i && \forall i \in \mathcal{I}_+, \\
& z_j - \lambda_j - \lambda_{[n]} = \sum_{i \in [n]: i \neq j} z_j^i + z_j^0 && \forall j \in [n], \qquad (8a)\\
& 0 \leq z_j^i \leq \lambda_i && \forall [n] \ni i \neq j \in [n] \cup \{0\}, \qquad (8b)\\
& \lambda \geq 0,\ \lambda_0 + \sum_{i=1}^{n} \lambda_i + \lambda_{[n]} = 1. && \qquad (8c)
\end{aligned}
$$

Next, we can substitute out $\sum_{i \in [n]: i \neq j} z_j^i + z_j^0$ in (8a) using the bounds (8b). Doing so, (8a) reduces to

$$
\begin{aligned}
& 0 \leq z_j - \lambda_j - \lambda_{[n]} \leq \sum_{i \in [n]: i \neq j} \lambda_i + \lambda_0 = 1 - \lambda_j - \lambda_{[n]}, \\
\Leftrightarrow & \lambda_{[n]} \leq z_j - \lambda_j,\ \ z_j \leq 1,
\end{aligned}
$$

where the equality results from (8c). We deduce that the system of inequalities reduces to

$$\begin{aligned} & t \geq \sum_{i=1}^{n} g^{\pi}(a_i(x_i - \tau_i), \lambda_i),\ a^{\top}\tau = 0, \\ & 0 \leq \tau_i \leq x_i && \forall i \in \mathcal{I}_+, \\ & z_j \leq 1,\ \lambda_{[n]} \leq z_j - \lambda_j && \forall j \in [n], \\ & \lambda \geq 0,\ \lambda_0 + \sum_{i=1}^{n} \lambda_i + \lambda_{[n]} = 1. \end{aligned}$$

Formulation (6) follows from using Fourier-Motzkin elimination to project out λ_0 and $\lambda_{[n]}$, replacing them with 0 and changing the last equality to an inequality. □

In addition, if $\mathcal{I}_+ = [n]$ and $a_i > 0\ \forall i \in [n]$, then $a^{\top}\tau = 0$ and $\tau \geq 0$ imply $\tau = 0$ in (6). Therefore, we deduce the following corollary for this special case.

Corollary 1. *Under Assumption 1, $\mathcal{I}_+ = [n]$ and $a_i > 0\ \forall i \in [n]$, $(t, x, z) \in$ cl conv$(\mathcal{Q})$ if and only if there exists $\lambda \in \mathbb{R}^n$ such that the inequality system*

$$\begin{aligned} & t \geq \sum_{i=1}^{n} g^{\pi}(a_i x_i, \lambda_i), \\ & \lambda_i \leq z_i \leq 1\ \forall i \in [n], \\ & \lambda \geq 0,\ \sum_{i=1}^{n} \lambda_i \leq 1 \end{aligned} \tag{10}$$

is satisfied.

Naturally, the extended formulations obtained from Proposition 3 and Corollary 1, requiring $\mathcal{O}(n)$ additional variables, are substantially more compact than extended formulations obtained from the disjunction (3), which require $\mathcal{O}(n^{2^n})$ variables.

3.2. **Explicit form of cl conv$(\mathcal{Q})$ with unconstrained continuous variables.** When x is unconstrained, i.e., $\mathcal{I}_+ = \emptyset$, the explicit form of cl conv$(\mathcal{Q})$ in the original space of variables was first established in [5] for quadratic functions, and later generalized in [38] to general rank-one functions. In Proposition 4 below, we present a short proof on how to recover the aforementioned results, starting from Proposition 3. First, we need the following property on the monotonicity of the perspective function.

Lemma 1. *Assume $g(v)$ is a convex function over $\mathbb{R}$ with $g(0) = 0$. Then $g^{\pi}(v, \lambda)$ is a nonincreasing function on $\lambda > 0$ for fixed v.*

Proof. Since g is convex, for $\forall v_1, v_2$, $\frac{g(v_1)-g(v_2)}{v_1-v_2}$ is nondecreasing with respect to v_1. Taking $v_1 = v/\lambda$ and $v_2 = 0$, since $g(0) = 0$, one can deduce that

$$g^\pi(v,\lambda) = \lambda g\left(\frac{v}{\lambda}\right) = v\frac{g\left(v\frac{1}{\lambda} - 0\right)}{v\frac{1}{\lambda} - 0}$$

is nondecreasing with respect to $1/\lambda$, i.e. nonincreasing with respect to λ. □

Proposition 4 (Wei et al. [38])**.** *Under Assumption 1, if additionally $\mathcal{I}_+ = \emptyset$, then*

$$cl\ conv(\mathcal{Q}) = \left\{(t,x,z) : t \geq g^\pi\left(a^\top x,\ \min\left\{1,\ \mathbf{e}^\top z\right\}\right),\ 0 \leq z \leq \mathbf{e}\right\}.$$

Proof. Without loss of generality, we assume $a = \mathbf{e}$; otherwise, we can scale x_i by a_i. We first eliminate τ from (6). From the convexity of $g(\cdot)$, we find that

$$\begin{aligned}
&\sum_{i\in[n]} g^\pi(x_i - \tau_i, \lambda_i) = \sum_{i\in[n]} \lambda_i g\left(\frac{x_i - \tau_i}{\lambda_i}\right) \\
&= \left(\sum_{j\in[n]} \lambda_j\right)\left(\sum_{i\in[n]} \frac{\lambda_i}{\sum_{j\in[n]}\lambda_j} g\left(\frac{x_i - \tau_i}{\lambda_i}\right)\right) \\
&\geq \left(\sum_{j\in[n]} \lambda_j\right) g\left(\frac{\sum_{i\in[n]} x_i - \sum_{i\in[n]}\tau_i}{\sum_{j\in[n]}\lambda_j}\right) \\
&= \left(\sum_{j\in[n]} \lambda_j\right) g\left(\frac{\sum_{i\in[n]} x_i}{\sum_{j\in[n]}\lambda_j}\right) = g^\pi\left(\sum_{i\in[n]} x_i, \sum_{i\in[n]}\lambda_i\right), \qquad (\mathbf{e}^\top\tau = 0)
\end{aligned}$$

where the inequality holds at equality if there exists some common ratio r such that $x_i - \tau_i = \lambda_i r$. Moreover, this ratio does exist by setting $r = \mathbf{e}^\top x/\mathbf{e}^\top\lambda$ and $\tau = x - r\lambda$ for all $i \in [n]$ –it can be verified directly that $\mathbf{e}^\top\tau = 0$. Thus, the above lower bound can be attained for all x. Hence, (6) reduces to

$$\begin{aligned}
&t \geq g^\pi(\mathbf{e}^\top x, \mathbf{e}^\top\lambda) \\
&0 \leq \lambda_i \leq z_i \leq 1\ \forall i \in [n] && \text{(11a)} \\
&\sum_{i\in[n]} \lambda_i \leq 1. && \text{(11b)}
\end{aligned}$$

Since for fixed v, $g^\pi(v,s)$ is non-increasing with respect to $s \in \mathbb{R}_+$, projecting out λ amounts to computing the maximum of $\mathbf{e}^\top\lambda$, that is, solving the linear program $\max_\lambda \{\mathbf{e}^\top\lambda :$ (11a) and (11b)$\}$. Summing up (11a) over all $i \in [n]$ and combining it with (11b), we deduce that $\mathbf{e}^\top\lambda \leq \min\{\mathbf{e}^\top z,\ 1\}$.

It remains to show this upper bound is tight. If $\sum_{i\in[n]} z_i \leq 1$, one can set $\lambda_i = z_i \ \forall i \in [n]$. Now assume $\sum_{i\in[n]} z_i > 1$. Let m be the index such that $\sum_{i=1}^{m-1} z_i \leq 1$ and $\sum_{i=1}^{m} z_i > 1$. Set

$$\lambda_i = \begin{cases} z_i & \text{if } i < m \\ 1 - \sum_{i=1}^{m} z_i & \text{if } i = m \\ 0 & \text{if } i > m. \end{cases}$$

It can be verified directly that this solution is feasible and $\mathbf{e}^\top \lambda = 1$. The conclusion follows. □

3.3. **Explicit form of cl conv($\mathcal{Q}$) with nonnegative continuous variables.** In this section, we aim to derive the explicit form of cl conv($\mathcal{Q}$) when $\mathcal{I}_+ = [n]$. A description of this set is known for the quadratic case [6] only. We now derive it for the general case. When specialized to bivariate rank-one functions, it also generalizes the main result of [4] where the research objective is a non-separable bivariate quadratic.

Observe that function f can be written as

$$f(x) = g\left(\sum_{i\in\mathcal{N}_+} a_i x_i - \sum_{i\in\mathcal{N}_-} a_i x_i\right),$$

where $a_i > 0 \ \forall i \in [n]$, and $\mathcal{N}_+ \cup \mathcal{N}_-$ is a partition of $[n]$. We first state the main theorem of this section, where the min / max over an empty set is taken to be $+\infty/-\infty$ respectively.

Theorem 2. *Under Assumption 1 and $\mathcal{I}_+ = [n]$, for all (t, x, z) such that $x \geq 0, 0 \leq z \leq \mathbf{e}$, the following statements hold:*

- *If $\sum_{i\in\mathcal{N}_+} a_i x_i > \sum_{i\in\mathcal{N}_-} a_i x_i$ and there exists a partition $\mathcal{L} \cup \mathcal{M} \cup \mathcal{U}$ of $\mathcal{N}_+ \cap \operatorname{supp}(x) \cap \operatorname{supp}(z)$ such that*

$$1 - \sum_{i\in\mathcal{M}\cup\mathcal{U}} z_i > 0, \quad \max_{i\in\mathcal{L}} \frac{a_i x_i}{z_i} < \frac{\sum_{i\in\mathcal{L}} a_i x_i}{1 - \sum_{i\in\mathcal{M}\cup\mathcal{U}} z_i} \leq \min_{i\in\mathcal{M}} \frac{a_i x_i}{z_i}$$

$$\max_{i\in\mathcal{M}} \frac{a_i x_i}{z_i} \leq \frac{\sum_{i\in\mathcal{N}_+} a_i x_i - \sum_{i\in\mathcal{L}\cup\mathcal{M}\cup\mathcal{N}_-} a_i x_i}{\sum_{i\in\mathcal{U}} z_i} < \min_{i\in\mathcal{U}} \frac{a_i x_i}{z_i},$$

then $(t, x, z) \in$ cl conv($\mathcal{Q}$) if and only if

$$t \geq g^\pi\left(\sum_{i\in\mathcal{L}} a_i x_i, 1 - \sum_{i\in\mathcal{M}\cup\mathcal{U}} z_i\right) + \sum_{i\in\mathcal{M}} g^\pi\left(a_i x_i, z_i\right) + g^\pi\left(\sum_{i\in\mathcal{N}_+} a_i x_i - \sum_{i\in\mathcal{L}\cup\mathcal{M}\cup\mathcal{N}_-} a_i x_i, \sum_{i\in\mathcal{U}} z_i\right). \quad (12)$$

Otherwise, $(t, x, z) \in$ cl conv($\mathcal{Q}$) if and only if $t \geq f(x)$.

- *If* $\sum_{i\in\mathcal{N}_+} a_ix_i \le \sum_{i\in\mathcal{N}_-} a_ix_i$ *and there exists an partition* $\mathcal{L}\cup\mathcal{M}\cup\mathcal{U}$ *of* $\mathcal{N}_-\cap\operatorname{supp}(x)\operatorname{supp}(z)$ *such that*

$$1-\sum_{i\in\mathcal{M}\cup\mathcal{U}} z_i > 0, \quad \max_{i\in\mathcal{L}} \frac{a_ix_i}{z_i} < \frac{\sum_{i\in\mathcal{L}} a_ix_i}{1-\sum_{i\in\mathcal{M}\cup\mathcal{U}} z_i} \le \min_{i\in\mathcal{M}} \frac{a_ix_i}{z_i}$$

$$\max_{i\in\mathcal{M}} \frac{a_ix_i}{z_i} \le \frac{\sum_{i\in\mathcal{N}_-} a_ix_i - \sum_{i\in\mathcal{L}\cup\mathcal{M}\cup\mathcal{N}_+} a_ix_i}{\sum_{i\in\mathcal{U}} z_i} < \min_{i\in\mathcal{U}} \frac{a_ix_i}{z_i},$$

then $(t,x,z)\in$ *cl conv*$(\mathcal{Q})$ *if and only if*

$$t \ge g^\pi\left(-\sum_{i\in\mathcal{L}} a_ix_i, 1-\sum_{i\in\mathcal{M}\cup\mathcal{U}} z_i\right) + \sum_{i\in\mathcal{M}} g^\pi\left(-a_ix_i, z_i\right) + g^\pi\left(-\sum_{i\in\mathcal{N}_-} a_ix_i + \sum_{i\in L\cup\mathcal{M}\cup\mathcal{N}_+} a_ix_i, \sum_{i\in\mathcal{U}} z_i\right).$$

Otherwise, $(t,x,z)\in$ *cl conv*$(\mathcal{Q})$ *if and only if* $t\ge f(x)$.

Note that $f(x)$ can be rewritten as $f(x)=\tilde{g}\left(\sum_{i\in\mathcal{N}_-} a_ix_i - \sum_{i\in\mathcal{N}_+} a_ix_i\right)$, where $\tilde{g}(s)=g(-s)$. Due to this symmetry, it suffices to prove the first assertion of Theorem 2. Without loss of generality, we also assume $a_i = 1\ \forall i\in[n]$; otherwise we can consider $\tilde{x}_i = a_ix_i$.

For simplicity, we first establish the convex hull results under the following assumption which is stronger than Assumption 1. Later, we will extend the conclusion of Theorem 2 to a general finite convex function.

Assumption 2. In addition to Assumption 1, function g is assumed to be a strongly convex and differentiable function with $g(0)=0=\min_{s\in\mathbb{R}} g(s)$. Also, we assume $x>0$, $0<z\le\mathbf{e}$ and $\sum_{i\in\mathcal{N}_+} a_ix_i > \sum_{i\in\mathcal{N}_-} a_ix_i$.

Observe that the positivity assumption $x>0$ and $z>0$ implies that $\mathcal{N}_+\cap\operatorname{supp}(x)\cap\operatorname{supp}(z)$ is simply $\mathcal{N}_+$. For any vector y and index set $\mathcal{S}\subseteq[n]$, we denote $y(\mathcal{S})=\sum_{i\in\mathcal{S}} y_i$ for convenience. The workhorse of the derivation of Theorem 2 is the following minimization problem induced by the extended formulation (6) of cl conv$(\mathcal{Q})$: for a given (x,z) such that $x>0, 0<z\le\mathbf{e}$ and $\sum_{i\in\mathcal{N}_+} x_i > \sum_{i\in\mathcal{N}_-} x_i$,

$$\begin{aligned} \min_{\lambda,\tau}\ & \sum_{i\in\mathcal{N}_+} \lambda_i g\left(\frac{x_i-\tau_i}{\lambda_i}\right) + \sum_{i\in\mathcal{N}_-} \lambda_i g\left(-\frac{x_i-\tau_i}{\lambda_i}\right) \\ \text{s.t. }\ & \tau(\mathcal{N}_+)-\tau(\mathcal{N}_-)=0,\ 0\le\tau\le x \\ & \mathbf{e}^\top\lambda\le 1,\ 0\le\lambda\le z. \end{aligned} \tag{13}$$

Lemma 2. *Under Assumption 2, one can set* $\tau_i=x_i,\ \lambda_i=0\ \forall i\in\mathcal{N}_-$ *and* $\lambda_i>0,\ \tau_i<x_i\ \forall i\in\mathcal{N}_+$ *in an optimal solution to* (13).

Proof. For contradiction, we assume there exists some $i_- \in \mathcal{N}_-$ such that $\tau_{i_-} < x_{i_-}$ in the optimal solution to (13). Since

$$\sum_{i\in\mathcal{N}_+} \tau_i = \sum_{i\in\mathcal{N}_-} \tau_i < \sum_{i\in\mathcal{N}_-} x_i < \sum_{i\in\mathcal{N}_+} x_i,$$

there must be some $i_+ \in \mathcal{N}_+$ such that $\tau_{i_+} < x_{i_+}$. Since g is convex and attains its minimum at 0, g is increasing over $[0, +\infty)$ and decreasing over $(-\infty, 0]$. It follows that increasing τ_{i_-} and τ_{i_+} by the same sufficiently small amount of $\epsilon > 0$ would improve the objective function, which contradicts with the optimality. Hence, $\tau_i = x_i \ \forall i \in \mathcal{N}_-$. It follows that one can safely take $\lambda_i = 0 \ \forall i \in \mathcal{N}_-$.

We now prove the second part of the statement, namely, that there exists an optimal solution with $\lambda > 0$. If $\lambda_i = 0, i \in \mathcal{N}_+$, since g is strongly convex, one must have $\tau_i = x_i$, otherwise $g^\pi(x_i - \tau_i, 0) = +\infty$. Moreover, if $\tau_i = x_i$, one can safely take $\lambda_i = 0$. Finally, assume $\lambda_i = 0, \tau_i = x_i$ for some index i. In this case, constraints $\tau(\mathcal{N}_+) - \tau(\mathcal{N}_-) = 0$, the previously proven property that $\tau(\mathcal{N}_-) = x(\mathcal{N}_-)$ and the assumption $x(\mathcal{N}_+) > x(\mathcal{N}_-)$ imply that there exists an index j where $\tau_j < x_j$ and $\lambda_j > 0$. Setting $(\tilde{\lambda}_i, x_i - \tilde{\tau}_i) = \epsilon(\lambda_j, x_j - \tau_j)$ and $(\tilde{\lambda}_j, x_j - \tilde{\tau}_j) = (1-\epsilon)(\lambda_j, x_j - \tau_j)$ for some small enough $\epsilon > 0$, we find that the new feasible solution is still optimal since $g^\pi(\cdot, \cdot)$ is a homogeneous function. The conclusion follows. □

By changing variable $\tau \leftarrow x - \tau$, it follows from Lemma 2 that (13) can be simplified to

$$\min_{\lambda,\tau} \sum_{i\in\mathcal{N}_+} \lambda_i g\left(\frac{\tau_i}{\lambda_i}\right) \tag{14a}$$

$$\text{s.t. } \tau(\mathcal{N}_+) = C \tag{14b}$$

$$0 < \tau \le x \tag{14c}$$

$$\lambda(\mathcal{N}_+) \le 1 \tag{14d}$$

$$0 < \lambda \le z, \tag{14e}$$

where $C \overset{\text{def}}{=} x(\mathcal{N}_+) - x(\mathcal{N}_-)$. Denote the derivative of $g(t)$ by $g'(t)$. Define $G(t) \overset{\text{def}}{=} tg'(t) - g(t) > 0$ and $\hat{g}(t) \overset{\text{def}}{=} tg(1/t)$. Next lemma will be used to simplify (14) and express the optimal solution to (14) in terms of $G(t)$ and $g'(t)$.

Lemma 3. *Under Assumption 2, $g'(t)$ and $G(t)$ are increasing over $(0, +\infty)$ and thus invertible. Moreover, $\hat{g}(t)$ is strictly convex over $(0, +\infty)$.*

Proof. Strict monotonicity of $g'(t)$ over $[0, +\infty)$ follows directly from the strict convexity of g. Since g is strongly convex, it can be written as $g(t) =$

$g_1(t) + g_2(t)$, where $g_1(t) = \alpha t^2$ is quadratic with a certain $\alpha > 0$ and $g_2(t)$ is convex. For any $t > 0$,

$$\begin{aligned}
&G(t+\epsilon) - G(t) = \epsilon g'(t+\epsilon) - (g(t+\epsilon) - g(t)) + t\left(g'(t+\epsilon) - g'(t)\right)\\
=&\epsilon\left(g'(t+\epsilon) - g'(t)\right) + t\left(g'(t+\epsilon) - g'(t)\right) + o(\epsilon) \qquad \text{(Taylor expansion)}\\
=&\epsilon\left(g'(t+\epsilon) - g'(t)\right) + t(g_2'(t+\epsilon) - g_2'(t)) + 2\alpha t\epsilon + o(\epsilon)\\
\geq&2\alpha t\epsilon + o(\epsilon) > 0, \quad \text{(Monotonicity of the derivative of a convex function)}
\end{aligned}$$

as ϵ is small enough. Namely, $G(t)$ is an increasing function over $[0, +\infty)$. Moreover, $G(0) = 0$ implies that $G(t) > 0\ \forall t \in (0, +\infty)$. To prove the last conclusion, $\hat{g}'(t) = g(1/t) - g'(1/t)/t = -G(1/t)$ is increasing over $(0, +\infty)$, which implies the strict convexity of $\hat{g}$. □

Assume (λ, τ) is an optimal solution to (14). Let $r_i \overset{\text{def}}{=} \frac{\tau_i}{\lambda_i}\ \forall i \in \mathcal{N}_+$. The next lemma reveals the structure of the optimal solution to (14). Namely, unless r_i's are identical, either λ_i or τ_i attains the upper bound.

Lemma 4. *Under Assumption 2, if $r_i > r_j$, then $\lambda_i = z_i$ and $\tau_j = x_j$ in the optimal solution to* (14)*.*

Proof. For contradiction we assume $\lambda_i < z_i$. Take $\epsilon > 0$ small enough and let $\epsilon_i = \epsilon/\tau_i$ and $\epsilon_j = \epsilon/\tau_j$. Since $\hat{g}$ is strictly convex and $1/r_i < 1/r_j$, one can deduce that

$$\begin{aligned}
&\frac{\hat{g}(1/r_i + \epsilon_i) - \hat{g}(1/r_i)}{\epsilon_i} < \frac{\hat{g}(1/r_j) - \hat{g}(1/r_j - \epsilon_j)}{\epsilon_j}\\
\Leftrightarrow\quad &\frac{(\lambda_i + \epsilon)/\tau_i g(\tau_i/(\lambda_i + \epsilon)) - \lambda_i/\tau_i g(\tau_i/\lambda_i)}{\epsilon/\tau_i} <\\
&\qquad\qquad \frac{\lambda_j/\tau_j g(\tau_j/\lambda_j) - (\lambda_j - \epsilon)/\tau_j g(\tau_j/(\lambda_j - \epsilon))}{\epsilon/\tau_j}\\
\Leftrightarrow\quad &(\lambda_i + \epsilon) g\left(\frac{\tau_i}{\lambda_i + \epsilon}\right) + (\lambda_j - \epsilon) g\left(\frac{\tau_j}{\lambda_j - \epsilon}\right) - \lambda_i g\left(\frac{\tau_i}{\lambda_i}\right) - \lambda_j g\left(\frac{\tau_j}{\lambda_j}\right) < 0.
\end{aligned}$$

It implies that we can improve the objective value of (14) by increasing λ_i and decreasing λ_j by ϵ, which contradicts with the optimality. Hence, one can deduce $\lambda_i = z_i$.

Similarly, the second conclusion follows by

$$\begin{aligned}
&\lambda_i g\left(\frac{\tau_i - \epsilon}{\lambda_i}\right) + \lambda_j g\left(\frac{\tau_j + \epsilon}{\lambda_j}\right) - \lambda_i g\left(\frac{\tau_i}{\lambda_i}\right) - \lambda_j g\left(\frac{\tau_j}{\lambda_j}\right)\\
=&(g'(r_j) - g'(r_i))\epsilon + o(\epsilon) < 0,
\end{aligned}$$

since $g'(\cdot)$ is strictly increasing. □

Finally, we are now ready to prove Theorem 2 under the additional Assumption 2.

Proposition 5. *Under Assumption 2, the conclusion in Theorem 2 holds true.*

Proof. We discuss two cases defined by $z(\mathcal{N}_+)$ separately.
Case 1: $z(\mathcal{N}_+) > 1$. Since the objective function of (14) is decreasing with respect to λ by Lemma 1, one must have $\lambda(\mathcal{N}_+) = 1$. Thus, program (14) can be reduced to

$$\begin{aligned} \min_{\lambda,\tau} \ & \sum_{i\in\mathcal{N}_+} \lambda_i g\left(\frac{\tau_i}{\lambda_i}\right) && \text{(15a)} \\ \text{s.t. } & \tau(\mathcal{N}_+) = C && (\alpha) \\ & 0 < \tau \le x && (\beta) \\ & \lambda(\mathcal{N}_+) = 1 && (\delta) \\ & 0 < \lambda \le z. && (\gamma) \end{aligned}$$

Assume (τ, λ) is the optimal solution to (15) and define $r_i = \tau_i/\lambda_i$. There are two possibilities – either all r_i's are identical or there are at least two distinct values of r_i's. In the former case, denote $r = r_i \ \forall i \in \mathcal{N}_+$, that is, $\tau_i = r\lambda_i \ \forall i \in \mathcal{N}_+$. Then $\tau(\mathcal{N}_+)/\lambda(\mathcal{N}_+) = r\lambda(\mathcal{N}_+)/\lambda(\mathcal{N}_+) = r$, and in particular $r = C$. In this case, (15a) reduces to $\lambda(\mathcal{N}_+)g(r) = g(C)$.

Now we assume there are at least two distinct values of r_i's. By Lemma 4, for all $i \in \mathcal{N}_+$, either $x_i = \tau_i$ or $\lambda_i = z_i$. It follows that $\mathcal{N}_+ = \mathcal{L} \cup \mathcal{M} \cup \mathcal{U}$, where

$$\mathcal{L} = \{i : \tau_i = x_i, \lambda_i < z_i\}, \ \mathcal{M} = \{i : \tau_i = x_i, \lambda_i = z_i\}, \ \mathcal{U} = \{\tau_i < x_i, \lambda_i = z_i\}. \tag{16}$$

Since all constraints of (15) are linear, KKT conditions are necessary and sufficient for optimality of (λ, τ). Let $(\alpha, \beta, \gamma, \delta)$ be the dual variables associated with each constraint of (15). It follows that $\gamma_i = 0 \ \forall i \in \mathcal{L}$ and $\beta_i = 0 \ \forall i \in \mathcal{U}$. The KKT conditions of (15) can be stated as follows (left

column: statement of the KKT condition; right column: equivalent simplification)

$$
\begin{array}{l|l|r}
g'\left(\frac{x_i}{\lambda_i}\right) - \alpha + \beta_i = 0 & \beta_i = \alpha - g'\left(\frac{x_i}{\lambda_i}\right) & \forall i \in \mathcal{L} \\
g'\left(\frac{x_i}{z_i}\right) - \alpha + \beta_i = 0 & \beta_i = \alpha - g'\left(\frac{x_i}{z_i}\right) & \forall i \in \mathcal{M} \\
g'\left(\frac{\tau_i}{z_i}\right) - \alpha = 0 & \tau_i = z_i (g')^{-1}(\alpha) & \forall i \in \mathcal{U} \\
g\left(\frac{x_i}{\lambda_i}\right) - \frac{x_i}{\lambda_i} g'\left(\frac{x_i}{\lambda_i}\right) + \delta = 0 & \lambda_i = \frac{x_i}{G^{-1}(\delta)} & \forall i \in \mathcal{L} \\
g\left(\frac{x_i}{z_i}\right) - \frac{x_i}{z_i} g'\left(\frac{x_i}{z_i}\right) + \delta + \gamma_i = 0 & \gamma_i = G\left(\frac{x_i}{z_i}\right) - \delta & \forall i \in \mathcal{M} \\
g'\left(\frac{\tau_i}{z_i}\right) - \frac{\tau_i}{z_i} g'\left(\frac{\tau_i}{z_i}\right) + \delta + \gamma_i = 0 & \gamma_i = G\left(\frac{\tau_i}{z_i}\right) - \delta & \forall i \in \mathcal{U} \\
\tau(\mathcal{U}) + x(\mathcal{M}) + x(\mathcal{L}) = C & \tau(\mathcal{U}) = C - x(\mathcal{M}) - x(\mathcal{L}) & \\
\lambda(\mathcal{L}) + z(\mathcal{M}) + z(\mathcal{U}) = 1 & \lambda(\mathcal{L}) = 1 - z(\mathcal{M}) - z(\mathcal{U}) &
\end{array}
$$

$0 < \tau_i < x_i \; \forall i \in \mathcal{U},\; 0 < \lambda_i < z_i \; \forall i \in \mathcal{L},\; \beta_i \geq 0 \; \forall i \in \mathcal{L} \cup \mathcal{M},\; \gamma_i \geq 0 \; \forall i \in \mathcal{M} \cup \mathcal{U}$.

Denote by $\bar{C} = C - x(\mathcal{M}) - x(\mathcal{L})$ and $\bar{z} = 1 - z(\mathcal{M}) - z(\mathcal{U})$. Then

$$
\begin{array}{lll}
\tau(\mathcal{U}) = z(\mathcal{U})(g')^{-1}(\alpha) = \bar{C} & \Rightarrow & \alpha = g'\left(\bar{C}/z(\mathcal{U})\right) \\
\lambda(\mathcal{L}) = x(\mathcal{L})/G^{-1}(\delta) = \bar{z} & \Rightarrow & \delta = G\left(x(\mathcal{L})/\bar{z}\right).
\end{array}
$$

Thus, one can first substitute out α and δ to get $\beta_i \; \forall i \in \mathcal{M}, \tau_i \; \forall i \in \mathcal{U}, \lambda_i \; \forall i \in \mathcal{L}, \gamma_i \; \forall i \in \mathcal{M}$. Then one can plug in $\lambda_i \; \forall i \in \mathcal{L}$ and $\tau_i \; \forall i \in \mathcal{U}$ to work out $\beta_i \; \forall i \in \mathcal{L}$ and $\gamma_i \; \forall i \in \mathcal{U}$. Hence, we deduce that the KKT system is equivalent to

$$
\begin{aligned}
&\beta_i = g'(\bar{C}/z(\mathcal{U})) - g'(x(\mathcal{L})/\bar{z}) \geq 0 && \forall i \in \mathcal{L} && \text{(17a)} \\
&\beta_i = g'(\bar{C}/z(\mathcal{U})) - g'(x_i/z_i) \geq 0 && \forall i \in \mathcal{M} && \text{(17b)} \\
&\tau_i = \bar{C} z_i / z(\mathcal{U}) \in (0, x_i) && \forall i \in \mathcal{U} && \text{(17c)} \\
&\lambda_i = \bar{z} x_i / x(\mathcal{L}) \in (0, z_i) && \forall i \in \mathcal{L} && \text{(17d)} \\
&\gamma_i = G(x_i/z_i) - G(x(\mathcal{L})/\bar{z}) \geq 0 && \forall i \in \mathcal{M} && \text{(17e)} \\
&\gamma_i = G(\bar{C}/z(\mathcal{U})) - G(x(\mathcal{L})/\bar{z}) \geq 0 && \forall i \in \mathcal{U}. && \text{(17f)}
\end{aligned}
$$

Because g' and G are increasing from Lemma 3, the KKT system has a solution if and only if

$$
\min_{i \in \mathcal{U}} \frac{x_i}{z_i} \overset{\text{(17c)}}{>} \frac{\bar{C}}{z(\mathcal{U})} \overset{\text{(17b)}}{\geq} \max_{i \in \mathcal{M}} \frac{x_i}{z_i} \geq \min_{i \in \mathcal{M}} \frac{x_i}{z_i} \overset{\text{(17e)}}{\geq} \frac{x(\mathcal{L})}{\bar{z}} \overset{\text{(17d)}}{>} \max_{i \in \mathcal{L}} \frac{x_i}{z_i},
$$

which implies $\frac{\bar{C}}{z(\mathcal{U})} \geq \frac{x(\mathcal{L})}{\bar{z}} \Leftrightarrow$(17a) and (17f). Moreover, by using the solution to the KKT system, the optimal value of (15) is

$$
g\left(\frac{x(\mathcal{L})}{\bar{z}}\right) + \sum_{i \in \mathcal{M}} z_i g\left(\frac{x_i}{z_i}\right) + z(\mathcal{U}) g\left(\frac{\bar{C}}{z(\mathcal{U})}\right).
$$

Case 2: $z(\mathcal{N}_+) \le 1$. Since the objective function of (14) is decreasing with respect to λ, one must have $\lambda_i = z_i \ \forall i \in \mathcal{N}_+$. Thus, problem (14) reduces to

$$\min_{\tau} \sum_{i\in\mathcal{N}_+} z_i g\left(\frac{\tau_i}{z_i}\right) \tag{18a}$$

$$\text{s.t. } \tau(\mathcal{N}_+) = C \tag{α}$$

$$0 < \tau \le x. \tag{β}$$

Assume τ is the optimal solution to (18). Similarly, define

$$\mathcal{L} = \emptyset, \quad \mathcal{M} = \{i : \tau_i = x_i\}, \quad \mathcal{U} = \{i : \tau_i < x_i\}.$$

Then the KKT conditions can be written as

$$\begin{array}{l|l|l}
g'\left(\frac{x_i}{z_i}\right) - \alpha + \beta_i = 0 & \beta_i = -g'\left(\frac{x_i}{z_i}\right) + \alpha & \forall i \in \mathcal{M} \\
g'\left(\frac{\tau_i}{z_i}\right) - \alpha & \tau_i = (g')^{-1}(\alpha) z_i & \forall i \in \mathcal{U} \\
x(\mathcal{M}) + \tau(\mathcal{U}) = C & \tau(\mathcal{U}) = C - x(\mathcal{M}) = \bar{C} = (g')^{-1}(\alpha) z(\mathcal{U}) &
\end{array}$$

$$0 < \tau_i < x_i \ \forall i \in \mathcal{U}, \ \beta_i \ge 0 \ \forall i \in \mathcal{M}.$$

It follows that $\alpha = g'(\bar{C}/z(\mathcal{U}))$. Plugging α in, one arrives at

$$\beta_i = g'(\bar{C}/z(\mathcal{U})) - g'(x_i/z_i) \ge 0 \qquad \forall i \in \mathcal{M}$$

$$\tau_i = z_i \bar{C}/z(\mathcal{U}) \in (0, x_i) \qquad \forall i \in \mathcal{U}.$$

Therefore, the KKT system has a solution if and only if

$$\min_{i\in\mathcal{U}} \frac{x_i}{z_i} > \frac{\bar{C}}{z(\mathcal{U})} \ge \max_{i\in\mathcal{M}} \frac{x_i}{z_i}.$$

The proof is finished. □

From Proposition 5, we see that Theorem 2 holds if additional assumptions are imposed – namely g is finite, strongly convex and differentiable, $0 = g(0) = \min_{t\in\mathbb{R}} g(t)$ and $x > 0, z > 0$. To complete the proof of the theorem, we now show how to remove the assumptions, one by one.

Proof of Theorem 2. Due to the symmetry mentioned above, we only prove the first conclusion in the theorem under Assumption 1 and $a = \mathbf{e}$. A key observation is that the optimal primal solution to (14), given by (17c) and (17d), does not involve function g and only relies on x and z (while the values of the dual variables does depend on g). Denote this optimal solution by (τ^*, λ^*) and the objective function of (14) by $h(\tau, \lambda; g)$. Then $h(\tau^*, \lambda^*; g) \le h(\tau, \lambda; g)$ for all feasible solutions (τ, λ) to (14) and all functions g satisfying Assumption 2.

If g is not a strongly convex function, one can consider $g_\epsilon(s) = g(s) + \epsilon s^2$, where $\epsilon > 0$. Since g_ϵ is strongly convex, the conclusion is applicable to g_ϵ.

One can deduce that for any feasible solution (τ, λ) of (14) $h(\tau^*, \lambda^*; g_\epsilon) \leq h(\tau, \lambda; g_\epsilon)$, which implies $h(\tau^*, \lambda^*; g) \leq h(\tau, \lambda; g)$ by letting $\epsilon \to 0$. Thus, the conclusion holds if g is a differentiable function with $0 = g(0) = \min_{t\in\mathbb{R}} g(t)$.

If g is not a differentiable function, one can consider its *Moreau-Yosida regularization* $e_\epsilon g(s) \stackrel{\text{def}}{=} \min_w \left\{g(w) + \frac{1}{2\epsilon}(s-w)^2\right\} \leq g(s)$, where $\epsilon > 0$. It follows that $e_\epsilon g(s)$ is a differentiable convex function; see Corollary 4.5.5, [30]. Moreover, $e_\epsilon g(s) \to g(s)$ as $\epsilon \to 0$; see Theorem 1.25, [36]. Because $h(\tau^*, \lambda^*; e_\epsilon g) \leq h(\tau, \lambda; e_\epsilon g)$, letting $\epsilon \to 0$, one can deduce that the conclusion holds for any finite convex function g with $0 = g(0) = \min_{t\in\mathbb{R}} g(t)$.

For a general finite convex function g with $g(0) = 0$, $\tilde{g}(s) = g(s) - cs$ is a convex function with $0 = \tilde{g}(0) = \min_{s\in\mathbb{R}} \tilde{g}(s)$, where $c \in \partial g(0)$. The conclusion follows by applying the theorem to $\tilde{g}$.

Finally, if there exists $i \in \mathcal{N}_+$ such that $x_i = 0$, then $\tau_i = 0$ in (13) which implies that one can safely set $\lambda_i = 0$ in (13). Hence, we can exclude the variables associated with index i from consideration and reduce the problem to a lower-dimensional case. For this reason, without loss of generality, we assume $x_i > 0\ \forall i \in [n]$. Define $\mathcal{N}_0 \stackrel{\text{def}}{=} \{i \in \mathcal{N}_+ : z_i = 0\}$, $\tilde{f}(x, z)$ as the RHS of (12), and $f^*(x, z)$ as the optimal value of (13). Let $r > 0$ be a sufficiently large number and consider z^r defined as $z_i^r = x_i/r$ if $i \in \mathcal{N}_0$ and $z_i^r = z_i$ otherwise. Note that $\lim_{r\to+\infty} z^r = z$. If there exists a partition $\mathcal{L} \cup \mathcal{M} \cup \mathcal{U}$ of $\mathcal{N}_+ \cap \operatorname{supp}(z)$ stated in the theorem, then $\mathcal{L} \cup \mathcal{M} \cup \tilde{\mathcal{U}}$ is the partition of $\mathcal{N}_+$ associated with (x, z^r) where $\tilde{\mathcal{U}} = \mathcal{U} \cup \mathcal{N}_0$. Since $z^r > 0$, the conclusion holds for (x, z^r), i.e. $f^*(x, z^r) = \tilde{f}(x, z^r)$. Because f^* and $\tilde{f}$ are closed convex functions, $f^*(x, z) = \lim_{r\to+\infty} f^*(x, z^r) = \lim_{r\to+\infty} \tilde{f}(x, z^r) = \tilde{f}(x, z)$. On the other hand, if for (x, z^r) such a partition $\mathcal{L} \cup \mathcal{M} \cup \mathcal{U}$ of $\mathcal{N}_+$ does not exist, then neither does for (x, z) because otherwise, $(\mathcal{L}\backslash\mathcal{N}_0, \mathcal{M}\backslash\mathcal{N}_0, \mathcal{U}\backslash\mathcal{N}_0)$ would be a proper partition of $\mathcal{N}_+ \cap \operatorname{supp}(z)$. In this case, the conclusion follows from the closedness of f and f^*. This completes the proof.

□

Remark 1. Sets $\mathcal{L}$, $\mathcal{M}$ and $\mathcal{U}$ in Theorem 2 can be found in $\mathcal{O}\left(n^2\right)$ time. Indeed, without loss of generality, we assume that $\sum_{i\in\mathcal{N}_+} a_i x_i > \sum_{i\in\mathcal{N}_-} a_i x_i$. First, sort and index x_i/z_i in a nondecreasing order. It follows from the conditions in Theorem 2 that if such $\mathcal{L}$, $\mathcal{M}$ and $\mathcal{U}$ exist, then there must be some k_1 and k_2 such that $\mathcal{L} = \{i \in \mathcal{N}_+ : i < k_1\}$, $\mathcal{M} = \{i \in \mathcal{N}_+ : k_1 \leq i \leq k_2\}$ and $\mathcal{U} = \{i \in \mathcal{N}_+ : i > k_2\}$. Consequently, one can verify the conditions in Theorem 2 by enumerating all possible combinations $\{k_1, k_2\} \subseteq \mathcal{N}_+$.

Now we turn to the special case where $\mathcal{N}_- = \emptyset$, that is, every entry of a is positive.

Corollary 2. *Under Assumption 1 and $a > 0$, point $(t, x, z) \in$ cl conv$(\mathcal{Q})$ if and only if $(x, z) \in [0,1]^n \times \mathbb{R}^n_+$ and there exists a partition $\mathcal{L} \cup \mathcal{M} =$ supp$(x) \cap$ supp(z) such that*

$$1 - \sum_{i\in\mathcal{M}} z_i \geq 0, \quad \max_{i\in\mathcal{L}} \frac{a_i x_i}{z_i} < \frac{\sum_{i\in\mathcal{L}} a_i x_i}{1 - \sum_{i\in\mathcal{M}} z_i} \leq \min_{i\in\mathcal{M}} \frac{a_i x_i}{z_i}, \tag{19}$$

and the following inequality holds

$$t \geq g^\pi\left(\sum_{i\in\mathcal{L}} a_i x_i, 1 - \sum_{i\in\mathcal{M}} z_i\right) + \sum_{i\in\mathcal{M}} g^\pi(a_i x_i, z_i) + g^\pi\left(\sum_{i\in\mathcal{N}_+} a_i x_i - \sum_{i\in\mathcal{L}\cup\mathcal{M}} a_i x_i, 0\right). \tag{20}$$

Proof. In this particular setting, it is easy to see that $\mathcal{N}_- = \emptyset$ and $\tau_i = 0\ \forall i \in [n]$ in (13). Thus, the partition defined in (16) always exists with $\mathcal{U} = \emptyset$. The conclusion follows from Theorem 2. □

Finally, we close this section by generalizing the main result of [4] to non-quadratic functions. Specifically, in [4], the authors studied the set

$$\mathcal{Q}_2 \stackrel{\text{def}}{=} \left\{(t, x, z) \in \mathbb{R}^3 \times \{0,2\}^n : \begin{array}{ll} t \geq g(a_1 x_1 - a_2 x_2), x_i \geq 0,\ i = 1,2, \\ x_i(1 - z_i) = 0,\ i = 1,2 \end{array}\right\},$$

where g is quadratic, and provided the description of cl conv$(\mathcal{Q}_2)$ in the original space of variable. A similar result holds for general convex functions.

Corollary 3. *Given a convex function $g(\cdot)$ with* dom$(g) = \mathbb{R}$ *and $f(x_1, x_2) = g(a_1 x_1 - a_2 x_2)$, where $a > 0$, point $(t, x, z) \in$ cl conv$(\mathcal{Q})$ if and only if $(x, z) \in [0,1]^2 \times \mathbb{R}^2_+$ and*

$$t \geq \begin{cases} g^\pi(a_1 x_1 - a_2 x_2, z_1) & \text{if } a_1 x_1 \geq a_2 x_2 \\ g^\pi(a_2 x_2 - a_1 x_1, z_2) & \text{if } a_2 x_2 \geq a_1 x_1. \end{cases}$$

Proof. In this case, $\mathcal{N}_+$ is a singleton in (14). □

3.4. **Implementation.** In this section, we discuss the implementation of the results given in Theorem 2 (for the quadratic case) with conic quadratic solvers.

A key difficulty towards using the convexification is that inequalities (12) are not valid: while they describe cl conv(Q) in their corresponding region, determined by partition $\mathcal{L} \cup \mathcal{M} \cup \mathcal{U}$, they may cut off points of cl conv(Q) elsewhere. To circumvent this issue, Atamtürk and Gómez [6] propose valid inequalities, each requiring $\mathcal{O}(n)$ additional variables and corresponding exactly with (12) in the corresponding region, and valid elsewhere. The inequalities are then implemented as cutting surfaces, added on the fly as needed. It is worth nothing that since the optimization problems considered

are nonlinear, and convex relaxations are solved via interior point solvers, adding a cut requires resolving again the convex relaxation (without the warm-starting capabilities of the simplex method for linear optimization).

In contrast, we can use Proposition 3 directly to implement the inequalities. When specialized to quadratic functions g, and with the introduction of auxiliary variables u to model conic quadratic cones, we can restate Proposition 3 as: $(x, y, t) \in \text{cl conv}(\mathcal{Q})$ if and only if there exists $(\lambda, \tau, u) \in \mathbb{R}^{3n}$ such that

$$t \geq \sum_{i=1}^{n} a_i^2 u_i, \tag{21a}$$

$$\lambda_i u_i \geq (x_i - \tau_i)^2,\ u_i \geq 0,\ \forall i \in [n], \tag{21b}$$

$$a^\top \tau = 0,\ 0 \leq \tau_i \leq x_i\ \forall i \in \mathcal{I}_+, \tag{21c}$$

$$\lambda_i \leq z_i \leq 1,\ \forall i \in [n], \tag{21d}$$

$$\lambda \geq 0,\ \sum_{i=1}^{n} \lambda_i \leq 1. \tag{21e}$$

is satisfied. Inequalities (21b) are (convex) rotated cone constraints, which can be handled by most off-the-shelf conic quadratic solvers, and every other constraint is linear. Note that using (21) requires adding $\mathcal{O}(n)$ variables *once* –instead of adding a similar number of variables *per inequality added*, with exponentially many inequalities required to describe cl conv$(\mathcal{Q})$–, and thus is a substantially more compact formulation than the one presented in [6].

To illustrate the benefits resulting from a more compact formulation, we compare the two formulations in instances used by [6], available online at `https://sites.google.com/usc.edu/gomez/data`. The instances correspond to portfolio optimization problems of the form

$$\min \sum_{k=1}^{K} t_k + \sum_{i=1}^{n} (d_i x_i)^2 \tag{22a}$$

$$\text{s.t. } t_k \geq (a_k^\top x)^2 \tag{22b}$$

$$\mathbf{e}^\top x = 1 \qquad \forall k \in [K] \tag{22c}$$

$$c^\top x - h^\top z \geq b \tag{22d}$$

$$0 \leq x \leq z \tag{22e}$$

$$x \in \mathbb{R}^n,\ z \in \{0,1\}^n, \tag{22f}$$

where $a_k \in \mathbb{R}^n$ for all $k \in [K]$, $c, d, h \in \mathbb{R}^n_+$ and $b \in \mathbb{R}_+$. Strong relaxations can be obtained by relaxing the integrality constraints $z \in \{0,1\}^n \to z \in [0,1]^n$, using the perspective reformulation $(d_i x_i)^2 \to (d_i x_i)^2 / z_i$ in the objective, and adding inequalities (21) (or using cutting surfaces) corresponding

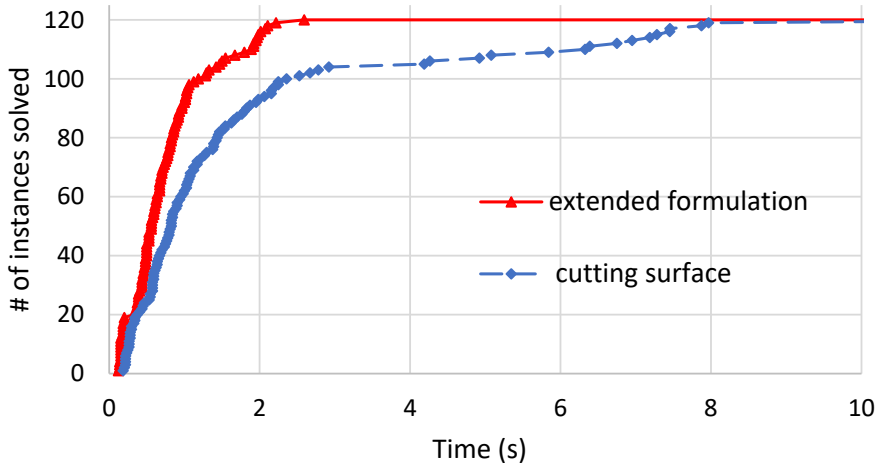

FIGURE 1. Number of instances solved as a function of time. The cutting surface requires on average 1.79 seconds to solve an instance, and can solve all 120 instances in 13.3 seconds or less. In contrast, the extended formulation (21) requires on average 0.78 seconds, and solves all instances in less than 2.6 seconds.

to each rank-one constraint (22b). Figure 1 summarizes the computational times required to solve the convex relaxations across 120 instances with $n \in \{200, 500\}$ and $K \in \{5, 10, 20\}$, using the CPLEX solver in a laptop with Intel Core i7-8550U CPU and 16GB memory. In short, the extended formulation (21) is on average twice as fast as the cutting surface method proposed in [6], and up to five times faster in the more difficult instances (in addition to arguably being easier to implement).

We also tested the effect of the extended formulation using CPLEX branch-and-bound solver. While we did not encounter numerical issues resulting in incorrect behavior by the solver (the cutting surface method does result in numerical issues, see [6]), the performance of the branch-and-bound method is substantially impaired when using the extended formulation. We discuss in more detail the issues of using the extended formulation with CPLEX branch-and-bound method in the appendix.

4. $\mathcal{NP}$-HARDNESS WITH BOUND CONSTRAINTS ON CONTINUOUS VARIABLES

The set $\mathcal{Q}$ studied so far assumes that the continuous variables are either unbounded, or non-negative/non-positive. Either way, set $\mathcal{Q}$ admits a similar disjunctive form given in Theorem 1 and, if it is rank-one, results

in similar compact extended formulations given in Proposition 3. A natural question is whether the addition of bounds on the continuous variables results in similar convexifications, or if the resulting set is structurally different. In this section, we show that it is impossible to describe the convex hull with bounded variables in a compact way, unless $\mathcal{P} = \mathcal{NP}$.

Consider the set

$$\mathcal{Q}_B \stackrel{\text{def}}{=} \left\{(t, x, z) \in \mathbb{R}^{n+1} \times \{0,1\}^n : t \geq f(x),\ 0 \leq x \leq z\right\}.$$

We show that describing cl conv($\mathcal{Q}_B$) is $\mathcal{NP}$-hard even when rank(f) = 1. Two examples are given to illustrate this point – *single node flow sets* and rank-one quadratic forms.

Single-node fixed-charge flow set. The single-node fixed-charge flow set is the mixed integer linear set defined as

$$\mathcal{T} \stackrel{\text{def}}{=} \left\{(x, z) \in \mathbb{R}^n \times \{0,1\}^n : \sum_{i=1}^n a_i x_i \leq b,\ 0 \leq x \leq z\right\},$$

where $0 < a_i \leq b\ \forall i \in [n]$. Note that one face of the single-node flow set conv($\mathcal{T} \cap \{(x, z) : x_i = z_i\ \forall i \in [n]\}$) is isomorphic to the knapsack set conv($\{z \in \{0,1\}^n : \sum_{i=1}^n a_i z_i \leq b\}$). If we define $f(x) = g\left(a^\top x\right)$, where $g(t) = \delta\left(t; \{t \in \mathbb{R} : t \leq b\}\right)$, it is clear that $\mathcal{Q}_B = \mathbb{R}_+ \times \mathcal{T}$, which means $\mathcal{T}$ is isomorphic to one facet of conv($\mathcal{Q}_B$). Thus, it is impossible to describe conv($\mathcal{Q}_B$) in a compact way unless $\mathcal{P} = \mathcal{NP}$.

Rank-one quadratic program with box-constrained continuous variables. Consider the following mixed-integer quadratic program

$$\begin{aligned} \min_{x,z}\ & \left(a^\top x\right)^2 + b^\top x + c^\top z \\ \text{s.t. }\ & 0 \leq x \leq z, \\ & z \in \{0,1\}^n. \end{aligned} \tag{23}$$

We aim to show (23) is $\mathcal{NP}$-hard in general. To achieve this goal, we show that (23) includes the following well known 0-1 knapsack problem (24) as a special case.

$$\begin{aligned} \min_{z \in \{0,1\}^n}\ & -v^\top z \\ \text{s.t. }\ & w^\top z \leq W, \end{aligned} \tag{24}$$

where $(v, w, W) \in \mathbb{Z}_+^{2n+1}$ are nonnegative integers such that $w_i \leq W \leq \sum_j w_j\ \forall i \in [n]$.

Proposition 6. *The knapsack problem* (24) *is equivalent to the optimization problem*

$$\begin{aligned} \min_{x,z}\ & M_1\left(Wx_0+\sum_{i=1}^n w_ix_i-W\right)^2-M_2\left(\sum_{i=1}^n x_i\right)+\sum_{i=1}^n(M_2-v_i)z_i \\ s.t.\ & 0\le x_i\le z_i,\ i=0,1,\dots,n \\ & z_i\in\{0,1\},\ i=0,1,\dots,n, \end{aligned} \tag{25}$$

where $M_1=\sum_{i=1}^n v_i+1$ *and* $M_2=2nW^2M_1+1$ *are polynomial in the input size.*

Proof. Denote the objective function by $\theta(x,z)$. First, since the objective function does not involve z_0, z_0 can be safely taken as 1. Second, we now show that M_2 is large enough to force $x_i=z_i$ for all $i\in[n]$. Specifically, for any $i\in[n]$, since $x_0,x_i\le 1$ and $w_i\le W$, it holds that

$$\begin{aligned} \frac{\partial\theta(x,z)}{\partial x_i} &= 2M_1w_i\left(Wx_0+\sum_{i=1}^n w_ix_i-W\right)-M_2 \\ &\le 2nW^2M_1-M_2<0. \end{aligned}$$

That is, $\theta(x,z)$ is decreasing with respect to $x_i, i\in[n]$, which implies $x_i=z_i$ in any optimal solution. It follows that (25) can be simplified to

$$\begin{aligned} \min_{x_0,z}\ & M_1\left(Wx_0+\sum_{i=1}^n w_iz_i-W\right)^2-\sum_{i=1}^n v_iz_i \\ \text{s.t.}\ & 0\le x_0\le 1,\ z_i\in\{0,1\},\ i=1,\dots,n, \end{aligned} \tag{26}$$

Next, we claim that M_1 is large enough to ensure that the optimal solution to (26) satisfies $Wx_0+\sum_{i=1}^n w_iz_i-W=0$, that is, $x_0=1-\left(w^\top z\right)/W$. To prove it rigorously, observe that the minimum value of (26) must be non-positive since $x_0=1, z=0$ is a feasible solution with objective value equal to 0. Moreover, since $1-\left(w^\top z\right)/W\le 1$, if $x_0\ne 1-\left(w^\top z\right)/W$ at the optimal solution to (26), then $1-\left(w^\top z\right)/W<0$ and the optimal x_0 must attain its lower bound 0. Furthermore, $1-\left(w^\top z\right)/W<0$ implies that $w^\top z\ge W+1$ since w, W and z are nonnegative integers. In this case, setting $x_0=0$, the minimum objective value of (26) can be written as

$$M_1\left(\sum_{i=1}^n w_iz_i-W\right)^2-\sum_{i=1}^n v_iz_i>M_1-\sum_i v_i=1>0,$$

which contradicts the non-positivity of the optimal objective value.

Therefore, we can substitute out $x_0 = 1 - \left(w^\top z\right)/W$ and (25) further reduces to

$$\begin{aligned} \min_{z} \ & -\sum_{i=1}^{n} v_i z_i \\ \text{s.t. } & 0 \leq 1 - \frac{w^\top z}{W} \leq 1, \\ & z_i \in \{0,1\},\ i = 1, \ldots, n, \end{aligned}$$

which is equivalent to (24), because $\left(w^\top z\right)/W \geq 0$ holds trivially. □

Due to the equivalence between optimization problems and separation problems [26], Proposition 6 indicates that it is impossible to extend the analysis in Section 3.1 and Section 3 to the case with bounded continuous variables.

5. Conclusions

In this paper, we propose a new disjunctive programming representation of the convex envelope of a low-rank convex function with indicator variables and complementary constraints. The ensuing formulations are substantially more compact than alternative disjunctive programming formulations. As a result, it is substantially easy to project out the additional variables to recover formulations in the original space of variables, and to implement the formulations using off-the-shelf solvers.

Acknowledgments

This research is supported in part by the National Science Foundation under grant CIF 2006762.

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*, 37:187–191.

[2] Anderson, R., Huchette, J., Tjandraatmadja, C., and Vielma, J. P. (2019). Strong mixed-integer programming formulations for trained neural networks. In *International Conference on Integer Programming and Combinatorial Optimization*, pages 27–42. Springer.

[3] Anstreicher, K. M. and Burer, S. (2021). Quadratic optimization with switching variables: the convex hull for $n = 2$. *Mathematical Programming*, 188(2):421–441.

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

[5] Atamturk, A. and Gomez, A. (2019). Rank-one convexification for sparse regression. *arXiv preprint arXiv:1901.10334*.

[6] Atamtürk, A. and Gómez, A. (2020). Supermodularity and valid inequalities for quadratic optimization with indicators. *arXiv preprint arXiv:2012.14633*.

[7] Balas, E. (1979). Disjunctive programming. *Annals of discrete mathematics*, 5:3–51.

[8] Balas, E. (1985). Disjunctive programming and a hierarchy of relaxations for discrete optimization problems. *SIAM Journal on Algebraic Discrete Methods*, 6(3):466–486.

[9] Balas, E. (1998). Disjunctive programming: Properties of the convex hull of feasible points. *Discrete Applied Mathematics*, 89(1-3):3–44.

[10] Balas, E. (2018). *Disjunctive programming*. Springer.

[11] Balas, E., Tama, J. M., and Tind, J. (1989). Sequential convexification in reverse convex and disjunctive programming. *Mathematical Programming*, 44(1):337–350.

[12] Bernal, D. E. and Grossmann, I. E. (2021). Convex mixed-integer nonlinear programs derived from generalized disjunctive programming using cones. *arXiv preprint arXiv:2109.09657*.

[13] Bertsimas, D., Cory-Wright, R., and Pauphilet, J. (2020). Mixed-projection conic optimization: A new paradigm for modeling rank constraints. *arXiv preprint arXiv:2009.10395*.

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

[15] Bienstock, D. (1996). Computational study of a family of mixed-integer quadratic programming problems. *Mathematical programming*, 74(2):121–140.

[16] Bonami, P. (2011). Lift-and-project cuts for mixed integer convex programs. In *International Conference on Integer Programming and Combinatorial Optimization*, pages 52–64. Springer.

[17] Ceria, S. and Soares, J. (1999). Convex programming for disjunctive convex optimization. *Mathematical Programming*, 86(3):595–614.

[18] Çezik, M. T. and Iyengar, G. (2005). Cuts for mixed 0-1 conic programming. *Mathematical Programming*, 104:179–202.

[19] Dantzig, G. B. (1972). Fourier-motzkin elimination and its dual. Technical report, STANFORD UNIV CA DEPT OF OPERATIONS RESEARCH.

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

[21] Dong, H. and Linderoth, J. (2013). On valid inequalities for quadratic programming with continuous variables and binary indicators. In Goemans, M. and Correa, J., editors, *Integer Programming and Combinatorial Optimization*, pages 169–180, Berlin, Heidelberg. Springer Berlin Heidelberg.

[22] Frangioni, A. and Gentile, C. (2006). Perspective cuts for a class of convex 0–1 mixed integer programs. *Mathematical Programming*, 106(2):225–236.

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

[24] Gómez, A. (2021). Strong formulations for conic quadratic optimization with indicator variables. *Mathematical Programming*, 188(1):193–226.

[25] Grossmann, I. E. (2002). Review of nonlinear mixed-integer and disjunctive programming techniques. *Optimization and engineering*, 3(3):227–252.

[26] Grötschel, M., Lovász, L., and Schrijver, A. (1981). The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1:169–197.

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

[28] Han, S. and Gómez, A. (2021). Single-neuron convexification for binarized neural networks. *http://www.optimization-online.org/DB_HTML/2021/05/8419.html*.

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

[30] Hiriart-Urruty, J.-B. and Lemaréchal, C. (2004). *Fundamentals of convex analysis*. Springer Science & Business Media.

[31] Kilinc, M., Linderoth, J., and Luedtke, J. (2010). Effective separation of disjunctive cuts for convex mixed integer nonlinear programs. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

[32] Kılınç-Karzan, F. and Yıldız, S. (2015). Two-term disjunctions on the second-order cone. *Mathematical Programming*, 154:463–491.

[33] Lee, S. and Grossmann, I. E. (2000). New algorithms for nonlinear generalized disjunctive programming. *Computers & Chemical Engineering*, 24(9-10):2125–2141.

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

[35] Rockafellar, R. T. (1970). Convex analysis.
[36] Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational analysis*, volume 317. Springer Science & Business Media.
[37] Stubbs, R. A. and Mehrotra, S. (1999). A branch-and-cut method for 0-1 mixed convex programming. *Mathematical programming*, 86(3):515–532.
[38] Wei, L., Gómez, A., and Kucukyavuz, S. (2020). Ideal formulations for constrained convex optimization problems with indicator variables. *arXiv preprint arXiv:2007.00107*.
[39] Wu, B., Sun, X., Li, D., and Zheng, X. (2017). Quadratic convex reformulations for semicontinuous quadratic programming. *SIAM Journal on Optimization*, 27:1531–1553.
[40] Yıldız, S. and Kılınç-Karzan, F. (2016). Low-complexity relaxations and convex hulls of disjunctions on the positive semidefinite cone and general regular cones. *Optimization Online*.

Appendix A. On computational experiments with branch-and-bound

As mentioned in §3.4, the extended formulation (21) did not produce good results when used in conjunction with CPLEX branch-and-bound solver. To illustrate this phenomenon, Table 1 shows details on the performance of the solver in a single representative instance, but similar behavior was observed in *all* instances tested. The table shows from left to right: the time required to solve the convex relaxation via interior point methods and the lower bound produced by this relaxation (note that this is not part of the branch-and-bound algorithm); the time required to process the root node of the branch-and-bound tree, and the corresponding lower bound obtained; the time used to process the branch-and-bound tree, the number of branch-and-bound nodes explored, and the lower bound found after processing the tree (we set a time limit of 10 minutes).

Table 1. Performance of CPLEX solver in an instance with $n = 500$ and $k = 10$. Default settings are used, and a time limit of 10 minutes is set. The optimal objective value in the particular instance is 1.47.

Method	Convex relaxation		Root node		Branch-and-bound		
	Time(s)	LB	Time(s)	LB	Time(s)	Nodes	LB
Without (21)	0.2	1.09	3.5	1.20	7.7	460	1.47
With (21)	2.4	1.30	45.9	0.90	600	4,073	0.99

Our expectations a priori were as follows: using inequalities (21) should result in harder convex relaxations solved, and thus less nodes explored

within a given time limit; on the other hand, due to improved relaxations and higher-quality lower bounds, the algorithm should be able to prove optimality after exploring substantially less nodes. Thus, there should be a tradeoff between the number of nodes to be explored and the time required to process each node. From Table 1, we see that there is no tradeoff in practice.

The performance of the solver without inequalities (21) is as expected. While just solving the convex relaxation via interior point methods requires 0.2 seconds, there is an overhead of 3 seconds to process the root node due to preprocessing/cuts/heuristics and additional methods used by the solver (and the quality of the lower bound at the root node is slightly improved as a result).Then, after an additional 4 seconds used to explore 460 nodes, optimality is proven.

The performance of the solver using inequalities (21) defied our expectations. In theory, the more difficult convex relaxation can be solved with an overhead of 2 seconds, resulting in a root improvement of $(1.30-1.09)/(1.47-1.09) = 55.3\%$ (better than the one achieved by default CPLEX). In practice, the overhead is 40 seconds, and results in a *degradation* of the relaxation, that is, the lower bound proved at the root node is worse than the natural convex relaxation of problem without inequalities (21). From that point out, the branch-and-bound progresses slowly due to the more difficult relaxations, and the lower bounds are worse throughout the tree. Even after the time limit of 10 minutes and over 4,000 nodes explored, the lower bound proved by the algorithm is still worse than the natural convex relaxation.

While we cannot be sure about the exact reason of this behavior, we now make an educated guess. Most conic quadratic branch-and-bound solvers such as CPLEX do not use interior point methods to solve relaxations at each node of the branch-and-bound tree, but rather rely on polyhedral outer approximations in an extended space to benefit from the warm-start capabilities of the simplex method. We conjecture that while formulation (21) might not be particularly challenging to solve via interior point methods, it might be difficult to construct a good-quality outer approximation of reasonable size. If so, then the actual relaxation used by the solver is possibly a poor-quality linear outer approximation of the feasible region induced by (21), and is still difficult to solve, resulting in the worst of both worlds. We point out that the second author has encountered similar counterintuitive behavior with solvers (other than CPLEX) based on linear outer approximations [4], suggesting that indeed this phenomenon stems from the implementation of outer-approximations in general (rather than a particular quirk specific to CPLEX).