5-2018

Convex Hulls, Relaxations, and Approximations of General Monomials and Multilinear Functions

Yibo Xu
Clemson University, simonorraynor@gmail.com

Convex Hulls, Relaxations, and Approximations of General Monomials and Multilinear Functions

---

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Mathematical Sciences

---

by
Yibo Xu
May 2018

Accepted by:
Dr. Warren Adams, Committee Chair
Dr. Xuhong Gao
Dr. Matthew Saltzman
Dr. Cole Smith

Abstract

Motivated by a variety of problems in global optimization and integer programming that involve multilinear expressions of discrete or continuous variables, this research derives approximations of multilinear functions, and studies the accuracy of these approximations through worst-case error-analyses:

- The derivation of the convex hull representations of large families of symmetric multilinear polynomials (SMPs) that are defined over box constraints through geometrical exploitation of the polytope symmetry and specially designed facet generation method; and
- The identification of the set of all points at which a nonnegative multilinear polynomial on a box vanishes, which applies to the identification of the set of all points which satisfy any facet at equality.
- The worst-case error analysis associated with linearizations of monomial expressions in bounded discrete and/or continuous variables: for certain families of variable-bound structures, the worst-case errors associated with convex hull forms are studied, along with the identification of all points which produce these errors.
- The worst-case error analysis associated with replacing the multilinear monomial term with a "best" approximating linear function, in contrast to the previous item on "convex hull linearization:" using the results of the first item, explicit convex hull forms are exploited to identify the "best" linear functions.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Linearization techniques are commonly used in both global optimization and integer programming, as many such problems involve multilinear expressions of discrete and/or continuous variables. This research derives new approximations of multilinear functions, compare these approximations to existing methods, and studies the accuracy of these approximations through worst-case error-analyses. Within the literature, such approximations exist in higher-dimensional spaces that employ auxiliary variables to replace nonlinear expressions as well as to most accurately represent the nonlinear functions. An ongoing challenge is to determine the worst-case errors associated with different approximations, and to obtain forms that provide minimal such errors. This effort has three main emphases, with each emphasis summarized below.

1.1 Convex Hull of Symmetric Multilinear Polynomial

The first emphasis is to derive convex hull representations of special symmetric multilinear polynomials (SMPs) that are defined over box constraints, and to determine those points which satisfy the facet-defining inequalities (facets) with equality. Specifically, we derive, through an explicit listing of all the facets, convex hull representations of the graph

$$G \equiv \{(\boldsymbol{x}, y) \in \mathbb{R}^n \times \mathbb{R} : \boldsymbol{x} \in X,\ y = m(\boldsymbol{x})\}, \tag{1.1}$$

where

$$X \equiv \{\boldsymbol{x} \in \mathbb{R}^n : l \leq x_j \leq u \; \forall \; j \in N\}, \tag{1.2}$$

and where

$$m(\boldsymbol{x}) = \sum_{i=2}^{n} c_i \left(\sum_{\substack{J \subseteq N \\ |J|=i}} \prod_{j \in J} x_j \right), \tag{1.3}$$

with l and u denoting, respectively, lower and upper bounds on the variables x_j, $l < u$. We also seek to identify the set of all points in G at which each facet is satisfied with equality. Here, and throughout the document, $N \equiv \{1, \ldots, n\}$. The function $m(\boldsymbol{x})$ is an SMP, with the "symmetry" describing the property that, given any $\boldsymbol{x} \in X$, all n-factorial permutations of this realization are also in X, and have the same functional value. No linear terms in $\boldsymbol{x}$ are found within $m(\boldsymbol{x})$ since the inclusion would alter the convex hull representation only by a linear adjustment to each facet.

Two important characteristics of this problem are as follows. First, the convex hull of G is known to be a polytope in $\mathbb{R}^{n+1}$ having 2^n extreme points $(\boldsymbol{x}, y)$ that correspond in a one-to-one manner with the extreme points of X so that the values of $\boldsymbol{x}$ are preserved and so that $y = m(\boldsymbol{x})$. Second, the convex hull representation of (1.1) depends upon the specific bounds l and u found in (1.2), and the specific function $m(\boldsymbol{x})$ found in (1.3); that is, the structure and number of facets vary greatly in terms of these parameters.

We make three contributions in this effort. First, we exploit the symmetry of the convex hull of G as a polytope in $\mathbb{R}^{n+1}$, and provide theoretical results that establish a valid inequality to be a facet, which applies to the convex hull generation for any SMP. Second, we derive all facets for three families of functions $m(\boldsymbol{x})$ and bounds l and u. For the first family, we present a new approach for deriving a known convex hull form when $m(\boldsymbol{x})$ is supermodular. For the second family, we use a similar approach to the first to derive the convex hull form when $m(\boldsymbol{x})$ is a monomial having $-l = u = 1$, and utilize this derivation for a more general type of SMP. For the last family, as a demonstration of the power of our approach, we analytically derive the convex hull form when $m(\boldsymbol{x})$ has a free parameter that makes it not supermodular, but still tractable. Third, we derive the set of points at which a nonnegative multilinear polynomial defined over a box vanishes; then for every facet derived for both families, we identify all points within G at which the facet is satisfied with equality.

We adopt a two-step approach for generating all the facets. The first step exploits the

problem structure to establish necessary conditions for a valid inequality to be eligible to be a facet. The second step again uses the problem structure, but this time to motivate and verify families of facets, and to invoke the necessary conditions to conclude that all such families have been obtained.

A listing of our contributions is given below.

1. Derive properties of facets, including necessary conditions for a valid inequality to be a facet.
2. Motivate and verify families of facets, and use the derived conditions to establish that all facets have been obtained.
3. Derive the set of points at which a nonnegative multilinear polynomial defined over a box vanishes.
4. Identify, for each facet, all points within G at which the facet is satisfied with equality.

Chapters 2 and 3 are dedicated to this emphasis.

1.2 Convex Hull Error

The second emphasis is to analyze the worst-case errors associated with linearizations of monomial expressions in bounded discrete and/or continuous variables. The process of linearization consists of replacing a monomial term with a continuous variable, and then defining linear inequalities to restrict, to the extent possible, that the newly-defined variable is equal to its intended product. The feasible points are those points for which the new variable is equal to the monomial. Different linearizations can be constructed, with the convex hull of feasible points affording the best approximation. Regardless of the approximation used, including that of the convex hull, the new variable is not forced to equal its intended product at all solutions, and thus approximation errors result. At any given point, the approximation error is defined to be the absolute difference between the value of the new variable and that of the approximated monomial. The worst-case error is the maximum such difference. Notably, the convex hull of feasible points is not polyhedral for general monomials, and its explicit form is known only for multilinear monomials with special variable bound structures. This effort studies, for large families of variable-bound structures, the worst-case errors associated with convex hull forms. An outline of this emphasis is as follows:

1. Derive convex hull representations of different families of monomial expressions, emphasizing those arising in 0-1 polynomial optimization.

2. Compute worst-case errors associated with the convex hull forms for various monomial expressions, including general monomials defined over the unit hypercube, monomials having identical positive lower and upper bounds, and monomials whose lower bounds are the negatives of the upper bounds.

Chapter 4 is dedicated to this emphasis.

1.3 Linear Approximation Error

The third emphasis, in contrast to the second effort on "convex hull linearization," analyzes the worst-case error associated with replacing the multilinear monomial term with a "best" approximating linear function. This approach is unique in that it introduces no auxiliary variables and/or constraints, and does not provide an outer-approximation. At any given point, the error with the linear function replacement is defined to be the difference between the functional value at that point and the desired multilinear monomial value, with the worst-case error maximum among all errors. Using the results of the second emphasis, we exploit the explicit convex hull forms to identify the best linear functions. Specific results are the following:

1. For various multilinear functions, identify the "best" approximating linear function which yields the minimum worst-case error.

2. Compute the worst-case errors associated with the best linear approximations, and the points at which the worst-case errors are realized, utilizing the results established from Chapter 3.

3. Demonstrate that best-linear function approximations are preferable to those associated with convex hulls for multilinear expressions in the sense that the former yields worst-case errors that are bounded above by the latter, with equality holding only for bilinear terms.

Chapter 5 is dedicated to this emphasis.

1.4 Background and Notation

Denote $N = \{1, \ldots, n\}$. For any integer $p > 0$, denote $[p] = \{0, 1, \ldots, p\}$; the vector of 1's is $\mathbb{1} \in \mathbb{R}^p$. The j^{th} unit vector in $\mathbb{R}^n$ is $\mathbf{e}_j$, and the vector of 0's as $\mathbf{0}$. conv($\bullet$) denotes the convex hull of the set $\bullet$. For a matrix A, its transpose is denoted by A^T.

Denote the set minus in the traditional sense: for set S and set T, $T \setminus S \equiv \{t \in T : t \notin S\}$.

A n-permuation $\sigma : N \to N$ is a map that is bijective. Here we use the conventional functional notation $i \mapsto \sigma(i)$ for all $i \in N$.

A function $g : \mathbb{R}^n \to \mathbb{R}$ is defined to be *supermodular* over $D \subseteq \mathbb{R}^n$ if $g(\boldsymbol{v}^1 \uparrow \boldsymbol{v}^2) + g(\boldsymbol{v}^1 \downarrow \boldsymbol{v}^2) \geq g(\boldsymbol{v}^1) + g(\boldsymbol{v}^2)$ for all $\boldsymbol{v}^1, \boldsymbol{v}^2 \in D$, where $(\boldsymbol{v}^1 \uparrow \boldsymbol{v}^2)$ and $(\boldsymbol{v}^1 \downarrow \boldsymbol{v}^2)$ denote the component-wise maximum and minimum vectors of $\boldsymbol{v}^1$ and $\boldsymbol{v}^2$, respectively.

The RLT is a general methodology for reformulating mixed-integer linear and polynomial programs for the purpose of obtaining tight linear programming relaxations. There is a rich body of literature on the topic (see, for example, [28, 29, 30]), but we restrict attention here to that of a box-constrained region of n variables x_j, where each x_j is restricted to lie between variable bounds L_j and U_j. Specifically, consider the box

$$X' \equiv [L_1, U_1] \times \ldots \times [L_n, U_n] \equiv \{\boldsymbol{x} \in \mathbb{R}^n : L_j \leq x_j \leq U_j \; \forall \; j \in N\}. \tag{1.4}$$

The RLT process that we apply to (1.4) consists of the two distinct steps of *reformulation* and *linearization*. The *reformulation* step computes products of the expressions $(x_j - L_j)$ and $(U_j - x_j)$, taken n at a time, such that one such expression is chosen for each j. In this manner, 2^n multilinear polynomial functions of degree n emerge. To elaborate, define the 2^n functions $F(K)$ so that

$$F(K) = \prod_{j \in K} (x_j - L_j) \prod_{j \in N \setminus K} (U_j - x_j) \text{ for each } K \subseteq N.$$

Then we have 2^n multilinear polynomial functions of the form

$$F(K) \; \forall \; K \subseteq N.$$

Each of these functions is nonnegative for all $\boldsymbol{x}$ in the box of (1.4), and the RLT enforces this nonnegativity to obtain the 2^n multilinear polynomial inequalities

$$F(K) \geq 0 \; \forall \; K \subseteq N, \tag{1.5}$$

that are satisfied for all $\boldsymbol{x}$ in the box.

The *linearization* step then substitutes a continuous variable w_J for each of the $2^n - (n+1)$

distinct product terms $\prod_{j\in J} x_j$ with $J \subseteq N$ and $|J| \geq 2$ that are found in (1.5). Denote the linearized form of each function $F(K)$ that is obtained via such substitutions as $f(K)$. The RLT then gives the following *polyhedral* set:

$$\left\{(\boldsymbol{x}, \boldsymbol{w}) \in \mathbb{R}^n \times \mathbb{R}^{2^n-(n+1)} : f(K) \geq 0 \; \forall \; K \subseteq N\right\}. \tag{1.6}$$

We adopt the notation that $\{\bullet\}_L$ is the linearized form of the vector $\bullet$ that is obtained by substituting $w_J = \prod_{j\in J} x_j$ for all $J \subseteq N$ with $|J| \geq 2$ throughout $\bullet$. In this manner, $f(K) = \{F(K)\}_L$ for all $K \subseteq N$.

In the algebraic sense, on X' with general lower and upper bounds, all the n variable multilinear polynomials actually form a vector space $\mathcal{M} = \text{span}\left\{\prod_{j\in J} x_j : J \subseteq N\right\}$ of dimension 2^n; here $\prod_{j\in\emptyset} x_j$ represents 1 as a polynomial. We know that $\{F(J) : J \subseteq N\}$ is also a basis, because

$$0 = \sum_{J\subseteq N} a_J \prod_{j\in J}(x_j - L_j) \prod_{j\in N\setminus J} (U_j - x_j)$$

implies that

$$0 = a_J \prod_{j=1}^{n}(U_j - L_j) \; \forall J,$$

when it is evaluated at vertex $\boldsymbol{E}_J$ having $x_j = \begin{cases} U_j & j \in J \\ L_j & j \in N \setminus J \end{cases}$.

Therefore any multilinear polynomial $p(\boldsymbol{x})$ can be expressed uniquely as

$$p(\boldsymbol{x}) = \sum_{J\subseteq N} a_J F(J),$$

and by the same evaluation at $\boldsymbol{E}_J$,

$$p(\boldsymbol{E}_J) = a_J \prod_{j=1}^{n}(U_j - L_j) \; \forall J,$$

and hence

$$p(\boldsymbol{x}) = \frac{1}{\prod_{j=1}^{n}(U_j - L_j)} \sum_{J\subseteq N} p(\boldsymbol{E}_J) F(J). \tag{1.7}$$

This means that such representation completely depends on the functional values of the polynomial

at the 2^n vertices of X'.

Now to see whether a polynomial inequality $p(\boldsymbol{x}) \geq 0$ is valid on X', we just need to see whether $p(\boldsymbol{E}_J) \geq 0$ for all J. The reverse is also true, as $F(J) \geq 0$ over X' for all J.

If we turn to the RLT polytope (1.6), as each inequality (facet) is constructed by substituting every $\prod_{j\in I} x_j, |I| \geq 2$ with w_I in $F(K) \geq 0$, the RLT polytope lies in the 2^n dimensional vector space — RLT-$(\boldsymbol{x}, \boldsymbol{w})$ space, or

$$\mathcal{R} = \text{span}\Big(\{w_J : J \subseteq N, |J| \geq 2\} \cup \{1, x_1, \ldots, x_n\}\Big).$$

This substitution (linearization $\{\bullet\}_L$) is an isomorphism $\phi : \mathcal{M} \to \mathcal{R}$, because it is a linear map and it is bijective. So its inverse map $\phi^{-1} : \mathcal{R} \to \mathcal{M}$ is tracing back from the w terms to the multilinear monomial terms:

$$\phi^{-1}\left(a_\emptyset + \sum_{j=1}^{n} a_{\{j\}} x_j + \sum_{J\subseteq N:|J|\geq 2} a_J w_J\right) = \sum_{J\subseteq N} a_J \prod_{j\in J} x_j.$$

This means that for any expression or inequality in the RLT-$(\boldsymbol{x}, \boldsymbol{w})$ space, through ϕ^{-1}, we can trace back to a multilinear polynomial expression or inequality, while all the coefficients are preserved.

Now we know $\{f(J) : J \subseteq N\}$ is a basis of $\mathcal{R}$, because of the fact that $\{F(J) : J \subseteq N\}$ is a basis of $\mathcal{M}$. So for a linear expression $r(\boldsymbol{x}, \boldsymbol{w})$ in the RLT-$(\boldsymbol{x}, \boldsymbol{w})$ space, it can be represented uniquely as

$$r(\boldsymbol{x}, \boldsymbol{w}) = \sum_{J\subseteq N} a_J f(J).$$

We denote further that $p(\boldsymbol{x}) \equiv \phi^{-1}\left(r(\boldsymbol{x}, \boldsymbol{w})\right)$ is the preimage of $r(\boldsymbol{x}, \boldsymbol{w})$ in $\mathcal{M}$, therefore

$$p(\boldsymbol{x}) = \phi^{-1}\left(r(\boldsymbol{x}, \boldsymbol{w})\right) = \sum_{J\subseteq N} a_J F(J).$$

By (1.7) we know the values of each a_J, and hence the exact representation of $r(\boldsymbol{x}, \boldsymbol{w})$ in terms of $f(J)$. The utility of this process is that a valid inequality $r(\boldsymbol{x}, \boldsymbol{w}) \geq 0$ for the RLT polytope of (1.6) relates to a unique nonnegative multilinear polynomial $p(\boldsymbol{x}) = \phi^{-1}\left(r(\boldsymbol{x}, \boldsymbol{w})\right)$ over X', and vice versa; the multipliers are in common when $r(\boldsymbol{x}, \boldsymbol{w}) \geq 0$ is represented as a nonnegative combination of $f(J)$ and when $p(\boldsymbol{x})$ is represented as a nonnegative combination of $F(J)$. Another way to put this, any inequality in the space $\mathcal{R}$ is valid for the RLT polytope (1.6), if and only if its correspon-

ding multilinear polynomial inequality in $\mathcal{M}$ is nonnegative, and if and only if the corresponding polynomial inequality is nonnegative at all vertices of X'.

Remark 1. *The utility of* (1.7) *is greater than expected, in the sense that it can be applied to show the following known facts:*

- *the convex hull of the graph of* $p(\boldsymbol{x})$ *on* X' *is polyhedral.*
- *the convex hull of* $\left\{(\boldsymbol{x}, \boldsymbol{w}) \in X' \times \mathbb{R}^{2^n-(n+1)} : w_I = \prod_{j\in I} x_j \ \forall\, I \subseteq N, |I| \geq 2\right\}$ *is* (1.6).

In fact, due to (1.7)*, for any multilinear* $p(\boldsymbol{x})$*,*

$$\begin{pmatrix} \boldsymbol{x} \\ p(\boldsymbol{x}) \end{pmatrix} = \sum_{J\subseteq N} \frac{F(J)}{\prod_{j=1}^n (U_j - L_j)} \begin{pmatrix} \boldsymbol{E}_J \\ p(\boldsymbol{E}_J) \end{pmatrix},$$

since it is true component-wise. Here, for a fixed $\boldsymbol{x} \in X'$*, the fractions serve as weights within a convex combination as they are clearly nonnegative and sum to* 1*. Therefore the above equation implies the first item. Moreover, for a fixed* $\boldsymbol{x}$*, these weights are shared by* all *multilinear polynomials. The same argument generalizes the above relation for* $\boldsymbol{p}(\boldsymbol{x})$*, an array of multilinear polynomials:*

$$\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{p}(\boldsymbol{x}) \end{pmatrix} = \sum_{J\subseteq N} \frac{F(J)}{\prod_{j=1}^n (U_j - L_j)} \begin{pmatrix} \boldsymbol{E}_J \\ \boldsymbol{p}(\boldsymbol{E}_J) \end{pmatrix},$$

which indicates that the convex hull of the "graph" of this array $\boldsymbol{p}(\boldsymbol{x})$ *only has* $\begin{pmatrix} \boldsymbol{E}_J \\ \boldsymbol{p}(\boldsymbol{E}_J) \end{pmatrix}$ *as the extreme points. More specifically, for* $\boldsymbol{w}(\boldsymbol{x})$ *where* $w_I(\boldsymbol{x}) = \prod_{j\in I} x_j$ *for all* $I \subseteq N, |I| \geq 2$ *:*

$$\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{w}(\boldsymbol{x}) \end{pmatrix} = \sum_{J\subseteq N} \frac{F(J)}{\prod_{j=1}^n (U_j - L_j)} \begin{pmatrix} \boldsymbol{E}_J \\ \boldsymbol{w}(\boldsymbol{E}_J) \end{pmatrix}.$$

Now, applying the linearization isomorphism ϕ *to the above, we have*

$$\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{w} \end{pmatrix} = \sum_{J\subseteq N} \frac{f(J)}{\prod_{j=1}^n (U_j - L_j)} \begin{pmatrix} \boldsymbol{E}_J \\ \boldsymbol{w}(\boldsymbol{E}_J) \end{pmatrix}.$$

For the second item, we just need to see that every $(\boldsymbol{x}, \boldsymbol{w})$ *within* (1.6) *is a convex combination of the*

points $(\boldsymbol{E}_J, \boldsymbol{w}(\boldsymbol{E}_J))$ *by the above equation: the fractional weights still sum to* 1 *and are nonnegative.*

1.4.1 Symmetric Polynomials and Symmetric Inequalities for $X = [l, u]^n$

For $X = [l, u]^n$, there is another way to look at the symmetric polynomials in $\mathcal{M}$ and symmetric expressions in $\mathcal{R}$. We know that all the vertices in X now have all entries equal to either l or u, therefore any symmetric polynomial $p(\boldsymbol{x})$ in $\mathcal{M}$ has identical $p(\boldsymbol{E}_J)$ values for a fixed $|J|$. Through the linearization isomorphism ϕ, we know that any symmetric expression in $\mathcal{R}$ has identical multipliers for $f(J)$ having a fixed $|J|$.

If we consider the symmetric polynomial subspace $\mathcal{S}$ of $\mathcal{M}$, by the Fundamental Theorem of Symmetric Polynomials, it is the *span* of the $(n + 1)$ elementary symmetric polynomials over n variables:

$$\sum_{J \subseteq N : |J| = k} \prod_{j \in J} x_j : k \in [n],$$

because any symmetric polynomial can be uniquely expressed as a polynomial of elementary symmetric polynomials, and $\mathcal{M}$ being a multilinear polynomial space implies that the expression as a polynomial of elementary symmetric polynomials has no powers nor multiplication of elementary symmetric polynomials. Based on the previous paragraph, we know that

$$\mathcal{S} = \text{span} \left\{ \sum_{J \subseteq N : |J| = k} \prod_{j \in J} x_j : k \in [n] \right\}$$

can be expressed in terms of $\left\{ \sum_{J \subseteq N : |J| = k} F(J) : k \in [n] \right\}$, so the latter is a basis of $\mathcal{S}$.

Chapter 2

Convex Hull Generation for Special SMPs

2.1 Introduction

Suppose a nonlinear differentiable function $g(\boldsymbol{x})$ having n variables appears in either the objective or the constraint of an optimization problem, and suppose the feasible region is a polytope $\Pi \subseteq \mathbb{R}^n$. Depending on the structure of Π, this problem is usually NP-hard: the optimum could occur at any vertex of Π, or any critical point in Π, or any critical point in the relative interior of any face of Π; see [9] for a multilinear g over the unit hypercube or in terms of 0-1 programming and [6] for a survey in 0-1 programming; see [7] for a quadratic g over a box and [11] for a survey in quadratic programming over box constraints. A common relaxation technique is to replace the function $g(\boldsymbol{x})$ by a continuous variable, say y, and implement linear auxiliary constraints in the $\boldsymbol{x}$ and y variables. These constraints are designed in a way that all points on the graph G of the function g, denoted as

$$G_g \equiv \{(\boldsymbol{x}, y) \in \Pi \times \mathbb{R} : y = g(\boldsymbol{x})\},$$

satisfy these constraints in the $(\boldsymbol{x}, y)$-space, and hence the convex hull of the graph, $\text{conv}\,(G_g)$, will satisfy all these constraints. The foregoing fact leads to two classes of relaxation: when the convex hull is a polyhedral set, the tightest finite set of linear constraints can be introduced, i.e., the Halfspace(H)-representation of the convex hull, which falls under the linearization technique;

when the convex hull is not a polyhedral set, there will not be a finite set of linear constraints that is the tightest relaxation of the convex hull, then we instead turn to nonlinear convex constraints to formulate the convex hull, which falls into the convexification technique. Relative to the massive amount of literature on the linearization technique, the convexification technique is less studied and very challenging; see [23]; so far the convex hull is still elusive for $g = x_1^a x_2^b$ on $\Pi = [0,1]^2$ where a, b are integers larger than 1. In certain optimization problems, especially when g appears only in the objective, only a convex envelope or concave envelope of g is needed instead of the convex hull. For such cases, we can always rewrite the problem as a maximization, so that we want the concave envelope of some function g (or $-g$). But it is still ideal to have a polyhedral concave envelope — a concave envelope with a polyhedral hypograph, so that the set of linear auxiliary constraints could be finite and tightest.

Rikun [25] analytically showed the following:

Theorem 2.1.1 (Theorem 1.1 in [25])**.** *The concave envelope of g is polyhedral if and only if it is constructed from only the vertices of* Π*, i.e.,*

$$\text{conv}\{(\boldsymbol{x}, y) \in \Pi \times \mathbb{R} : y \leq g(\boldsymbol{x})\} = \text{conv}\{(\boldsymbol{x}, y) \in \Pi \times \mathbb{R} : y \leq g(\boldsymbol{x}), \boldsymbol{x} \text{ a vertex of } \Pi\}.$$

This means, if the concave envelope of g on Π differs from the concave envelope constructed from the vertices of Π at even one point, then the concave envelope of g is not polyhedral. A similar statement is true for $\text{conv}\,(G_g)$, i.e., if the convex hull is polyhedral, then it must be of the form:

$$\text{conv}\{(\boldsymbol{x}, y) \in \Pi \times \mathbb{R} : y = g(\boldsymbol{x})\} = \text{conv}\{(\boldsymbol{x}, y) \in \Pi \times \mathbb{R} : y = g(\boldsymbol{x}), \boldsymbol{x} \text{ a vertex of } \Pi\}.$$

On arbitrary box constraints $\Pi = \prod_{j=1}^{n} [L_j, U_j]$, Rikun [25] showed that the convex hull of a multilinear function is polyhedral. Using reformulation-linearization-technique (RLT) and Kronecker product, Xu et al. [35] alternatively shows the same result and clearly points out the mathematical relation: since the convex hull in the RLT extended space is constructed from the vertices of box Π, through a projection down to the $(\boldsymbol{x}, y)$-space, the same result holds. Moreover, $\text{conv}\{(\boldsymbol{x}, y) \in \Pi \times \mathbb{R} : y = y^{\boldsymbol{x}}, \boldsymbol{x} \text{ a vertex of } \Pi\}$, where $y^{\boldsymbol{x}}$ denotes the y-coordinate of the extreme point at vertex $\boldsymbol{x}$, is the convex hull of a unique multilinear polynomial. These facts clearly indicate the rationale of studying the convex hull of multilinear polynomials on box constraints; namely, that for any

differentiable function g defined on Π, its polyhedral convex hull will coincide with the convex hull of some multilinear polynomial. Even if its convex hull is not polyhedral, as long as the desired convex or concave envelope is polyhedral, it coincides with that of some multilinear polynomial.

The related approaches in the literature are summarized below:

- McCormick [20], Meyer and Floudas [21] and Meyer [22] gave the convex hull for a multilinear function having $n \leq 3$ on an arbitrary box in $\mathbb{R}^n$.

- Glover and Woolsey [18] and Crama [10] gave the convex hull for the multilinear monomial $\prod_{j=1}^{n} x_j$ on the unit hypercube.

- Tardella [32] introduced *edge concavity* — g being concave along each direction in Π that parallels to an edge of Π — as a sufficient condition for the convex envelope to be polyhedral.

- Del Pia and Khajavirad [15] studies the "multilinear set," which is constructed from linearizing all relevant monomial terms that have appeared in the expression of g simultaneously. This approach relates to the RLT approach via a projection.

- Benson [4] gave the *concave envelope* for the multilinear monomial $\prod_{j=1}^{n} x_j$ on the nonnegative box $\prod_{j=1}^{n}[L_j, U_j]$ with $L_j \geq 0$ for all j, through Kuhn's triangulation, and affine interpolation on each triangle.

- Sherali [27] gave the convex hull for basic SMPs, i.e., $g(\boldsymbol{x}) = \sum_{J:|J|=i} \prod_{j\in J} x_j$ for all $i \in N$, on the unit hypercube, by converting the concave envelope into a Linear Programming (LP) problem

$$\begin{aligned} \mathcal{P}(\boldsymbol{x}) \equiv \min \quad & \boldsymbol{\beta}^T \boldsymbol{x} + \beta_0 \\ \text{s.t.} \quad & \boldsymbol{\beta}^T \boldsymbol{v} + \beta_0 \geq g(\boldsymbol{v}) \ \forall \boldsymbol{v} \text{ a vertex of } \Pi \\ & \boldsymbol{\beta} \in \mathbb{R}^n, \ \beta_0 \in \mathbb{R} \end{aligned}$$

  and uses the primal-dual pair to conclude that each facet relates to a vertex of the primal feasible region

$$R_1 \equiv \{(\boldsymbol{\beta}, \beta_0) : \boldsymbol{\beta}^T \boldsymbol{v} + \beta_0 \geq g(\boldsymbol{v}) \ \forall \boldsymbol{v} \text{ a vertex of } \Pi\}. \tag{2.1}$$

  This approach relies on the enumeration of all the vertices of the above region R_1.

- Tawarmalani et al. [33] extended the approach of [27] by relating each facet to an element of a polyhedral subdivision — a collection of polytopes whose union is the grand polytope and whose interiors have empty pairwise intersections — of Π, so that the concave envelope can be computed by the affine interpolation on each element of the subdivision. Tawarmalani et al. [33] generalized [4] and [27] by giving the concave envelope for a multilinear polynomial g, which is supermodular when restricted to the vertices of the unit hypercube, and gave the convex envelope for an SMP g on the unit hypercube. This approach relies solely on the existing subdivisions of Π that guarantee polyhedral envelopes.
- Adams et al. [1] gave the convex hull form for the monomial $\prod_{j=1}^{n} x_j$ on the box $[-1, 1]^n$.
- Xu et al. [35] exploited the problem symmetry algebraically through defining the "core" family of facets that motivates all remaining facets. Necessary and sufficient conditions for "core" facets are obtained and utilized to derive the convex hull forms for a supermodular SMP on the box $[l, u]^n$ and the monomial $\prod_{j=1}^{n} x_j$ on the box $[-1, 1]^n$.

The results near the bottom of the list are highly dependent on the symmetry of the multilinear function g and its domain $\Pi = [l, u]^n$. The structure of the convex hull for an SMP is very special because the set of extreme points only rely on the vertices of the domain box and their functional values, and every vertex with a fixed number of entries at upper bound u shares a same functional value. So essentially, we are constructing a polytope of the following form

$$\text{conv}\{(\boldsymbol{x}, y) \in \Pi \times \mathbb{R} : y = y^k, \boldsymbol{x} \text{ a vertex having } k \text{ entries of } u\}, \tag{2.2}$$

which is symmetric with respect to $\boldsymbol{x}$. Instead of trying to enumerate the extreme points from the associating R_1 of (2.1) or identify the correct domain subdivision, as in [33], we follow along the approach from [35] of exploiting the symmetry of the problem and generating the facets of the convex hull directly. We focus on the convex hull generation of SMPs: given

$$G \equiv \{(\boldsymbol{x}, y) \in X \times \mathbb{R} : y = m(\boldsymbol{x})\},$$

where

$$m(\boldsymbol{x}) = \sum_{i=2}^{n} c_i \left(\sum_{\substack{J \subseteq N \\ |J|=i}} \prod_{j \in J} x_j \right)$$

is an SMP, and $X \equiv [l, u]^n$, the H-representation of conv (G) is desired. As motivated in [35], the polytope conv (G) is of the form (2.2) and symmetric with respect to $\boldsymbol{x}$, and its H-representation is highly redundant — a single facet can be permuted and produce a huge number of facets. A "core" collection of facets is then of interest. In Section 2.2, we introduce discrete functions, to characterize the polytope (2.2) and the "core" facets, and translate the symmetry exploitation along with the necessary and sufficient conditions for "core" facets from [35] graphically. Moreover, in Section 2.2.2, we complete the necessary and sufficient condition for a "core" valid inequality to be a facet; in Section 2.3, we provide a graphical argument which simplifies the convex hull generation of the special SMPs from [35]; in Section 2.3.2, we use the graphical insights to extend the result of the monomial $\prod_{j=1}^{n} x_j$ on $[-1, 1]^n$ for SMPs of a more general form.

With sufficient theoretical development from Section 2.2, there are two approaches for the "core" facet generation: analytical and numerical. In Section 2.4.1, we discuss the analytical procedure, which extends the graphical argument appearing in Section 2.3 and exploits the relatively simple problem structure, and we generate an analytical convex hull form for an SMP that is not supermodular (nor submodular). When a general problem structure takes place, we then must utilize the necessary and sufficient condition for "core" facets from Section 2.2.2 to get a complete classification of "core" facet types, which is discussed in [34].

2.2 Characteristics of Facets

In this section, we follow along the path of symmetry exploitation and reduction, which leads to the "core" family of facets from [35], aiming at the necessary and sufficient condition for a valid inequality to be a facet. For completeness of the results, we will refer to some important results from [35] along the path.

Before we get deeply involved with symmetry, we list two basic observations of the dimension of conv(G) and its facets:

Observation 2.2.1. *The dimension of* conv(G_g) *is either* n *or* $(n+1)$. *It is* n *if and only if* $g(\boldsymbol{x})$

is an affine function on X.

Proof. Firstly, the projection of $\text{conv}(G_g)$ down to the $\boldsymbol{x}$-space is X, which is of dimension n. This means that the dimension of $\text{conv}(G_g)$ is at least n. Moreover, if the dimension of $\text{conv}(G_g)$ is indeed n, then it is contained in a hyperplane in the $(\boldsymbol{x}, y)$-space. This hyperplane must have a y term, otherwise its projection down to the $\boldsymbol{x}$-space is of dimension $(n-1)$, which contradicts the fact that the subset $\text{conv}(G_g)$ has a dimension n projected down to the $\boldsymbol{x}$-space.

Now we assume that $\text{conv}(G_g)$ is contained in the hyperplane $y = \boldsymbol{\beta}^T\boldsymbol{x} + \beta_0$, then this hyperplane contains G_g as well, which means that $g(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x} + \beta_0$ on X. □

It is not at all interesting to have an affine $g(\boldsymbol{x})$, because then $\text{conv}(G_g) = G_g$, and the facet-defining inequalities (facets) are of dimension $(n-1)$, not even hyperplanes of the $(n+1)$ dimensional $(\boldsymbol{x}, y)$-space. However, if $g(\boldsymbol{x})$ is affine, then in an optimization problem, there is no necessity to linearize $g(\boldsymbol{x})$, as both its concave and convex envelope is itself. In other words, when $\text{conv}(G_g)$ is n dimensional, both the convex and concave envelope facets are the affine hull of $\text{conv}(G_g)$ itself. As we care about the concave and convex envelope in the optimization context, we are aiming at components of the envelopes of the $(n+1)$ dimensional $(\boldsymbol{x}, y)$-space, so the facets for the rest of the paper we consider are the n dimensional hyperplanes of the said space, and are all *non-vertical* facets which have y term in the inequalities.

Observation 2.2.2. *Let $y = \boldsymbol{\beta}^T\boldsymbol{x} + \beta_0$ be a non-vertical hyperplane in the $(n+1)$-dimensional $(\boldsymbol{x}, y)$-space. Any collection of p points lie on the hyperplane is affinely independent if and only if it is a collection of p affinely independent points only in terms of $\boldsymbol{x}$.*

Proof. For any collection of p points on the hyperplane, say $\begin{pmatrix} \boldsymbol{v}^k \\ y^k \end{pmatrix}$ for $k = 1, \ldots, p$, one just need to observe that:

$$\begin{bmatrix} 1 & \ldots & 1 \\ \boldsymbol{v}^1 & \ldots & \boldsymbol{v}^p \\ y^1 & \ldots & y^p \end{bmatrix} = \begin{bmatrix} 1 & \ldots & 1 \\ \boldsymbol{v}^1 & \ldots & \boldsymbol{v}^p \\ \boldsymbol{\beta}^T\boldsymbol{v}^1 + \beta_0 & \ldots & \boldsymbol{\beta}^T\boldsymbol{v}^p + \beta_0 \end{bmatrix}$$

and

$$\begin{bmatrix} 1 & \ldots & 1 \\ \boldsymbol{v}^1 & \ldots & \boldsymbol{v}^p \end{bmatrix}$$

have the same row rank, and hence the same matrix rank. □

This observation tells us that, to argue $(n+1)$ affinely independent points on a facet, one just needs to focus on the $\boldsymbol{x}$ portion of these points.

2.2.1 Symmetric Multilinear Polynomial and "Symmetric" Polytope

In order to construct $\text{conv}(G)$, we find it equivalent to construct a "symmetric" polytope in general. Our process of convex hull generation can be viewed as the process of facet generation for a "symmetric" polytope. This subsection serves as the starting point of our study of polytope construction, and as a preparation for introducing "Convex Covers" and "Plots."

As proved in [25] and [35], we know that the convex hull, $\text{conv}(G_g)$, of a multilinear polynomial g in the $(\boldsymbol{x}, y)$-space, is essentially a polytope of 2^n extreme points which are generated from the vertices of the domain box. As shown in [35], this correspondence is one-to-one from the set of multilinear polynomials to the set of 2^n extreme points. Moreover, if the polynomial itself is symmetric in $\boldsymbol{x}$ and has a symmetric domain X, i.e., if any permutation among the entries of a input vector $\boldsymbol{x}$ results in the same output value, the convex hull is symmetric in $\boldsymbol{x}$ as well, i.e., the convex hull contains $(\boldsymbol{x}, y)$ if and only if it contains $(\boldsymbol{x}_\sigma, y)$ for all n-permutation σ, where $\boldsymbol{x}_\sigma \equiv (x_{\sigma(1)}, \ldots, x_{\sigma(n)})^T$. (If a set has some symmetry property, then the convex hull of that set preserves that property; the proof uses a point-set argument upon the application of Carathéodory's Theorem from [8], which we omit here.)

The symmetry of such polytope can be characterized in terms of the 2^n extreme points only. We split the vertices of X into $(n+1)$ *levels*: a vertex of X is level-k if it has k entries of u and $(n-k)$ entries of l. We split the extreme points of the polytope in the same fashion: an extreme point is level-k if its $\boldsymbol{x}$-component is level-k. Then the polytope is symmetric in $\boldsymbol{x}$ if and only if its 2^n extreme points satisfy that, for each $k \in [n]$, all level-k extreme points have the same y value. (Here the "only if" direction is trivial, the "if" direction again uses Carathéodory's Theorem.) Moreover, if we define a multilinear polynomial in a similar way, that all level-k vertices of X give the same output value, of course this leads to an SMP. (This is the consequence of the fact that the collection of all SMPs as a vector spaces has the following two bases: $\left\{\sum_{J\subseteq N:|J|=k} \prod_{j\in J} x_j : k \in [n]\right\}$ and $\left\{\sum_{J\subseteq N:|J|=k} \prod_{j\in J}(x_j - l) \prod_{j\in N\setminus J}(u - x_j) : k \in [n]\right\}$.)

Summarizing the foregoing discussion, we state the following observation:

Observation 2.2.3. *1. For a multilinear polynomial $g(\boldsymbol{x})$ on X, the following are equivalent:*

(a) $g(\boldsymbol{x})$ is symmetric,

(b) $\text{conv}(G_g)$ is symmetric in $\boldsymbol{x}$,

(c) the level-k extreme points of $\text{conv}(G_g)$ have the same y value, for each $k \in [n]$.

2. Let P be a polytope of 2^n extreme points associated with the vertices of X, such that extreme points from the same level are assigned to a same y value, then there is a unique multilinear polynomial $g(\boldsymbol{x})$ which satisfies that $P = \text{conv}(G_g)$ (and such $g(\boldsymbol{x})$ is symmetric).

3. For a symmetric function $g(\boldsymbol{x})$ (not necessarily polynomial), $\text{conv}(G_g)$ is symmetric in $\boldsymbol{x}$, and the level-k extreme points of $\text{conv}(G)$ have the same y value, for each $k \in [n]$.

We see from the discussion that, if we want to study $\text{conv}(G)$ for all SMP $m(\boldsymbol{x})$, we can alternatively study all the polytopes which are constructed from the 2^n vertices of X in such a way that vertices from the same level are assigned to a same y value. Meanwhile, even with a symmetric function $g(\boldsymbol{x})$ which is not a polynomial, as long as we are willing to assume that the convex or concave envelope of G_g is polyhedral (with exactly 2^n extreme points, each associates with a vertex of X, by Theorem 2.1.1,) we are still studying such a polytope. Let P be such a polytope, i.e., for some $(y^0, \ldots, y^n) \in \mathbb{R}^{n+1}$,

$$P = \text{conv}\left\{(\boldsymbol{x}, y) : y = y^k, \text{ for level-}k \text{ vertex } \boldsymbol{x} \in X,\ k \in [n]\right\}. \tag{2.3}$$

We know P is a symmetric polytope in $\boldsymbol{x}$. According to Observation 2.2.1, if we want P to be full dimensional, the bottom line is that $n \geq 2$.

Observation 2.2.4. *1. $y = \boldsymbol{\beta}^T\boldsymbol{x} + \beta_0$ defines a facet of P if and only if $y = \boldsymbol{\beta}_\sigma^T\boldsymbol{x} + \beta_0$ defines a facet of P for any n-permutation σ.*

2. Similarly, $\beta' y + \boldsymbol{\beta}^T\boldsymbol{x} + \beta_0 \geq 0$ is a valid inequality (facet) of P if and only if $\beta' y + \boldsymbol{\beta}_\sigma^T\boldsymbol{x} + \beta_0 \geq 0$ is a valid inequality (facet) of P for any n-permutation σ, where $\beta' \in \{\pm 1\}$. Hence we focus on the "core" valid inequalities (facets) having $\beta_1 \leq \beta_2 \leq \ldots \leq \beta_n$, and motivate all facets by permuting all core facets.

3. *P can be completely summarized by $(y^0, \ldots, y^n)$. This is a one-to-one correspondence between the set of all such polytopes and $\mathbb{R}^{n+1}$. Moreover, P is $(n+1)$ dimensional if and only if, in 2D, the set of points $\{(k, y^k) : k \in [n]\}$ is not on one line.*

4. *Given any SMP $m(\boldsymbol{x})$, the corresponding y^k are determined by $y^k = m(\boldsymbol{E}_k)$ for $k \in [n]$, where $\boldsymbol{E}_k = \sum_{j=1}^{k} u\boldsymbol{e}_j + \sum_{j=k+1}^{n} l\boldsymbol{e}_j$, the vector with first k entries equal to u, last $(n-k)$ entries equal to l.*

5. *Given any symmetric polytope P summarized as $(y^0, \ldots, y^n) \in \mathbb{R}^{n+1}$, the corresponding SMP $m(\boldsymbol{x})$ which has P as the convex hull, is given by*

$$m(\boldsymbol{x}) = (u-l)^{-n} \sum_{k=0}^{n} y^k \sum_{J:|J|=k} \prod_{j \in J} (x_j - l) \prod_{j \in N \setminus J} (u - x_j).$$

Item 1 and 2 are the standard symmetry exploration as in [35]. Item 5 utilizes the study on the vector space of multilinear polynomials from [35], to provide the relation from a symmetric polytope P to its associating SMP.

To lay the foundation for the graphical treatment and the main results, we borrow the concept of *discrete* function with an additional requirement, so that we can define *restriction* unconventionally.

Definition 1. *A function $\mathcal{D}$ is a discrete function if it is defined on $[t]$ for some nonnegative integer t. The "restriction" of $\mathcal{D}$ on $A = \{i, \ldots, j\} \subseteq [t]$, a nonempty set of successive integers, is the discrete function $\mathcal{D}|_A(\bullet) = \mathcal{D}(\bullet + i)$ defined on $[j - i]$.*

As an example, P can be summarized as a discrete function, which we call a *plot* (for polytope P). The restriction is extracting a successive portion of a discrete function, and we treat it as a new discrete function, which will be further discussed in the next subsection.

Definition 2. *A discrete function is a plot if it is defined on $[t]$ for some positive integer t. The plot for the polytope P of (2.3) is defined as $\mathcal{P}(k) = y^k$ for all $k \in [n]$.*

Any plot $\mathcal{D}$ on $[t]$ defined in terms of Definition 2 is reasonable, because it is the plot for the polytope $\operatorname{conv}\{(\boldsymbol{x}, y) : y = \mathcal{D}(k)$, for level-$k$ vertex $\boldsymbol{x} \in [l, u]^t$, $k \in [t]\}$. In this sense, a restriction of a plot defined from Definition 1 is also a reasonable plot as long as the cardinality of A is at least 2.

In actuality we do not care how the plot and its restriction relate algebraically, because we only worry about getting a sufficiently sized set of affinely independent points for a facet, but the following remark relates a plot to its restriction.

Remark 2. *Let $\mathcal{P}$ defined on $[n]$ be the plot of symmetric polytope P, and have restriction $\mathcal{P}|_A$, where $A = \{i, \ldots, j\}$, $0 \leq i < j \leq n$. Then the symmetric polytope associated to $\mathcal{P}|_A$ in the $(j - i + 1)$ dimensional space is equal to the polytope produced from cross sectioning P by letting $x_1 = \ldots = x_{i-1} = u, x_{j+1} = \ldots = x_n = l$. The "$\subseteq$" direction is trivial, however the simplest proof of the "$\supseteq$" direction again uses a point-set argument upon applying Carathéodory's Theorem.*

2.2.2 Facets of Symmetric Polytope

In this subsection, we give characteristics of facets for the proposed symmetric polytopes. We motivate and introduce convex covers over plots — the graphical translation of core facets for SMPs, and investigate the characteristics of facets.

Notice that we only need to consider the core facets as described in item 2 of Observation 2.2.4. The very essence of symmetry exploitation, as established in [35], is to decrease the number of points needed to be checked for certifying the validity of a core inequality, from 2^n to $(n + 1)$. As the idea of the proof plays a central role in later applications, we present it here.

Proposition 2.2.1. *For the polytope P, given that $\beta_1 \leq \beta_2 \leq \ldots \leq \beta_n$ and $\beta' \in \{\pm 1\}$, the validity of $\beta' y + \boldsymbol{\beta}^T \boldsymbol{x} + \beta_0 \geq 0$ is equivalent to $\beta' y^k + \boldsymbol{\beta}^T \boldsymbol{E}_k + \beta_0 \geq 0 \, \forall k \in [n]$, where $\boldsymbol{E}_k = \sum_{j=1}^{k} u\boldsymbol{e}_j + \sum_{j=k+1}^{n} l\boldsymbol{e}_j$.*

Proof. For the validity of $\beta' y + \boldsymbol{\beta}^T \boldsymbol{x} + \beta_0 \geq 0$, we just need to guarantee that the inequality is true for all extreme points of P. Once we plug in all the level-k vertices of X, i.e., $\boldsymbol{x} = \sum_{j \in J} u\boldsymbol{e}_j + \sum_{j \in N \setminus J} l\boldsymbol{e}_j, \forall J \subseteq N, |J| = k$ for a fixed integer k, we notice that the left-hand side is minimized at $\boldsymbol{x} = \sum_{j=1}^{k} u\boldsymbol{e}_j + \sum_{j=k+1}^{n} l\boldsymbol{e}_j = \boldsymbol{E}_k$. Thus validity is equivalent to $\beta' y^k + \boldsymbol{\beta}^T \boldsymbol{E}_k + \beta_0 \geq 0$ holding true for all $k \in [n]$. □

$\boldsymbol{E}_k, \forall k \in [n]$, as identified in [35], are the vertices of the simplex

$$\{\boldsymbol{x} \in \mathbb{R}^n : u \geq x_1 \geq \ldots \geq x_n \geq l\},$$

which relates to Khun's triangulation. Triangulation is used in [4] and [33] for functions which have polyhedral concave envelopes and are supermodular when restricted to the vertices, relying on the fact that on each triangle, the affine interpolation hyperplane is valid. In general, however, a non-vertical facet takes its effect on a small polytope in X, with a "positive volume", but not necessarily a simplex.

For this inequality to be facet defining, by Observation 2.2.2, we wish that the equality holds at as many vertices of X as possible, so that there are $(n+1)$ affinely independent points in X. Therefore, we need the equality to hold at as many values of k as possible, and we need to intelligently choose collection(s) of entries in $\boldsymbol{\beta}$ to equal each other in the case that the equality holds at not all $k \in [n]$.

Notice that $\mathcal{P}(k) = y^k \, \forall k \in [n]$ is already defined as a discrete function in terms of Definition 2. To exploit the assumption on the core valid inequality that $\beta_1 \leq \beta_2 \leq \ldots \leq \beta_n$, we instead separately treat $\beta' \mathcal{P}(k) + \boldsymbol{\beta}^T \boldsymbol{E}_k + \beta_0 \geq 0$ as $\boldsymbol{\beta}^T \boldsymbol{E}_k + \beta_0 \geq \mathcal{P}(k)$ for the over-estimator of polytope P, and $\boldsymbol{\beta}^T \boldsymbol{E}_k + \beta_0 \geq -\mathcal{P}(k)$ for the under-estimator of polytope P, and define the left-hand side as a discrete function with the argument k as well:

$$\mathcal{C}_{\boldsymbol{\beta},\beta_0}(k) \equiv \boldsymbol{\beta}^T \boldsymbol{E}_k + \beta_0 = \sum_{j=1}^{k} u\beta_j + \sum_{j=k+1}^{n} l\beta_j + \beta_0 \, \forall k \in [n], \tag{2.4}$$

as the same algebraic summary was done in [35].

So $\boldsymbol{\beta}^T \boldsymbol{x} + \beta_0$ is an over-estimator of polytope P if and only if $\mathcal{C}_{\boldsymbol{\beta},\beta_0}(k) \geq \mathcal{P}(k) \, \forall k$; $-\left(\boldsymbol{\beta}^T \boldsymbol{x} + \beta_0\right)$ is an under-estimator of polytope P if and only if $\mathcal{C}_{\boldsymbol{\beta},\beta_0}(k) \geq -\mathcal{P}(k) \, \forall k$. Notice also that $-\mathcal{P}$ is actually the plot for the polytope $-P$, the polytope flipped from P in y; this means that we can focus on the over-estimator of polytope P, disregard the under-estimator for the moment, and then turn to the over-estimator of $-P$ instead. Thus this is simply enforcing that $\mathcal{C}_{\boldsymbol{\beta},\beta_0}(\bullet)$ as a discrete function is lower bounded by the plot of a polytope, which is easy to handle on a 2D graph.

Proposition 2.2.2. *$\mathcal{C}_{\boldsymbol{\beta},\beta_0}(\bullet)$ has nondecreasing first forward differences.*

Proof. By the definition in (2.4),

$$\mathcal{C}_{\boldsymbol{\beta},\beta_0}(k) - \mathcal{C}_{\boldsymbol{\beta},\beta_0}(k-1) = (u-l)\beta_k \; \forall k = 1, \ldots, n,$$

which is a non-decreasing difference as k increases. $\square$

The above property of $\mathcal{C}(\bullet)$ is shared by all core valid inequalities and facets (under the assumption from item 2 of Observation 2.2.4), and hence should also be enforced. Although it is defined as a discrete function, it is very helpful to illustrate it as a piecewise linear function constructed by connecting neighbor points, because on a 2D graph the first forward differences can be seen as the slopes of each piece, so the above criteria is requiring $\mathcal{C}_{\boldsymbol{\beta},\beta_0}(\bullet)$ to have nondecreasing slopes, in other words, to be convex, as a piecewise linear function.

Definition 3. *A discrete function $\mathcal{D}$ is called* convex *if it has nondecreasing first forward differences. It is called* linear *if it has constant first forward differences. Similarly, it is called* concave *if it has nonincreasing first forward differences.*

Moreover, we can abbreviate the discrete function notation $\mathcal{C}_{\boldsymbol{\beta},\beta_0}(\bullet)$ as $\mathcal{C}(\bullet)$ when $(\boldsymbol{\beta},\beta_0)$ is not used in the context. This is because for every convex function $\mathcal{C}(\bullet)$ on $[n]$, we can easily calculate its associating $\boldsymbol{\beta},\beta_0$ via the following nonsingular linear transformation:

$$\beta_k = \frac{\mathcal{C}(k)-\mathcal{C}(k-1)}{u-l}\ \forall k=1,\dots,n;\ \beta_0 = \frac{u\mathcal{C}(0)}{u-l} - \frac{l\mathcal{C}(n)}{u-l}. \tag{2.5}$$

Hence each such $\mathcal{C}(\bullet)$ is associated with a unique core over-estimating hyperplane for P.

Now that the problem of finding a (facet defining) valid inequality has become the problem of finding a (facet inducing) convex discrete function $\mathcal{C}(\bullet)$ which "covers" $\mathcal{P}(\bullet)$, we study the convex cover diagram, on the plot $\mathcal{P}(\bullet)$, all the possible ways to cover the plot from above using a convex function. By item 3 of Observation 2.2.4, the only assumption on $\mathcal{P}$ is that $\mathcal{P}$ cannot be linear, so that $\mathcal{P}$ is reasonable to be a plot for some symmetric polytope of full dimension.

Definition 4. *A* convex cover $\mathcal{C}$ *on plot $\mathcal{P}$ is a discrete function defined on the same domain as $\mathcal{P}$, such that:*

1. $\mathcal{C}(\bullet) \geq \mathcal{P}(\bullet)$,
2. $\mathcal{C}(\bullet)$ *has nondecreasing first forward differences, i.e., it is convex.*

We define and focus on the convex covers of $\mathcal{P}$, i.e., we are focusing on the core over-estimators of polytope P. We did not explicitly specify the domain on which a convex cover is defined, for the convenience of performing restrictions; the same is done for the definition below.

Definition 5. *A convex cover $\mathcal{C}$ on plot $\mathcal{P}$ is* facet inducing*, or* F.I.*, if its associating inequality defines a facet of the concave envelope of the polytope corresponding to $\mathcal{P}$.*

To be more specific, we need $(n+1)$ affinely independent extreme points of P on the hyperplane, but as Observation 2.2.2 suggests, it is equivalent for us to consider $(n+1)$ affinely independent points, within the Cover-Plot diagram, among the vertices of X.

Remark 3. *As opposed to the algebraic treatment in [35], we are formulating and translating the problem in a geometric manner. Either approach essentially relates to the facet enumeration approaches by [27] and [33].*

Before we provide the necessary and sufficient conditions for a convex cover to be F.I., we introduce a few concepts for a convex cover.

Definition 6. *Let $\mathcal{C} = \mathcal{C}_{\boldsymbol{\beta},\beta_0}$ be a convex cover on plot $\mathcal{P}$, which are both defined on $[n]$. Let P denote the symmetric polytope that associates to $\mathcal{P}$.*

1. *$T(\mathcal{C})$ is the set of places where $\mathcal{C}$* touches *$\mathcal{P}$, i.e., $T(\mathcal{C}) = \{k : \mathcal{C}(k) = \mathcal{P}(k)\}$;*
2. *$M(\mathcal{C})$ is the set of points at which the hyperplane defined by $\mathcal{C}$* meets *the extreme points of P, i.e., $M(\mathcal{C}) = \bigcup_{k=0}^{n} \{\boldsymbol{x} \in X \text{ level-}k \text{ vertex} : \mathcal{P}(k) = \boldsymbol{\beta}^T\boldsymbol{x} + \beta_0\}$;*
3. *The* pattern *of $\mathcal{C}$ is a collection $\{\{j \in N : \beta_j = \beta_k\} : k \in N\}$ in which we omit the repeated sets.*

Remark 4. *If we want $\mathcal{C}$ to be F.I., by definition, we precisely need $(n+1)$ affinely independent vertices in $M(\mathcal{C})$. To compute $M(\mathcal{C})$, we need both the structure of $\boldsymbol{\beta}$ and the places where $\mathcal{C}$ touches $\mathcal{P}$.*

1. *The pattern of $\mathcal{C}$ is a partition of N; so there are 2^{n-1} possible different patterns in general.*
2. *The pattern of $\mathcal{C}$ and $T(\mathcal{C})$ gives $M(\mathcal{C})$. The reverse is not true when $\mathcal{C}$ is not F.I. In this situation, $M(\mathcal{C})$ is shared by a collection of facets, and hence by a convex combination of them, there is no unique pattern that can be recovered.*
3. *However, if $\mathcal{C}$ induces a facet, then $M(\mathcal{C})$ gives both the pattern of $\mathcal{C}$ and $T(\mathcal{C})$. This is because the associating $\boldsymbol{\beta}^T\boldsymbol{x} + \beta_0$ can be calculated explicitly.*

The following condition is the necessary and sufficient condition for a linear cover to be F.I., and is part of the necessary and sufficient condition for a convex cover to be F.I.

Proposition 2.2.3. *Let $n \geq 2$. A linear cover $\mathcal{C}$, i.e., a convex cover that is also linear, on plot $\mathcal{P}$ is F.I. if and only if $T(\mathcal{C}) = \{k : \mathcal{C}(k) = \mathcal{P}(k)\} \neq \{0, n\}$ and has cardinality greater than or equal to* 2.

Proof. We just need to argue that, in X, the set of level-k vertices uniquely defines the $(n-1)$ dimensional hyperplane $\sum_{j=1}^{n} x_j = ku + (n-k)l$ for each $k \neq 0, n$. Equivalently, the following $(n+1) \times \binom{n}{k}$ matrix, formed by the $p = \binom{n}{k}$ vertices, has rank n,

$$\begin{bmatrix} 1 & \dots & 1 \\ \boldsymbol{v}^1 & \dots & \boldsymbol{v}^p \end{bmatrix}.$$

Notice that through subtracting l or u times the first row from the remaining rows, one only needs to argue the above statement for $1 \leq k \leq n/2$. And it is also sufficient to show the above statement only when X is the unit hypercube, i.e., $l = 0$ and $u = 1$.

Now as the first row is depending on the rest of the rows, it is equivalent to show that there is a collection of n vertices, such that the following $n \times n$ matrix

$$\begin{bmatrix} \boldsymbol{v}^1 & \dots & \boldsymbol{v}^n \end{bmatrix}$$

has full rank.

Let the following $n \times n$ matrix represent the n vertices of choice,

$$\left[\begin{array}{c|c} K & O \\ \hline L & I \end{array}\right]$$

where K is the $(k-1) \times (k-1)$ matrix with 0's along the diagonal, but 1's elsewhere; O is the $(k-1) \times (n-k+1)$ matrix with all 1's, L is the $(n-k+1) \times (k-1)$ matrix with bottom two rows all 1's, but 0's elsewhere, and I is the $(n-k+1) \times (n-k+1)$ identity matrix. So, each row has k 1's and 0's elsewhere, and it is easy to show its invertibility by performing column and row operations on its determinant. □

It is natural to do a comparison for each concept we have defined among convex covers.

Thus we borrow the terms "finer" and "coarser" from partition comparison for pattern comparison.

Definition 7. *Let* $\mathcal{C}^1 = \mathcal{C}_{\hat{\boldsymbol{\beta}},\hat{\beta}_0}$ *and* $\mathcal{C}^2 = \mathcal{C}_{\boldsymbol{\beta},\beta_0}$ *be convex covers on plot* $\mathcal{P}$*, defined on* $[n]$*.*

1. *If* $M(\mathcal{C}^1) \supseteq M(\mathcal{C}^2)$*, we say* $\mathcal{C}^1$ *is* superior *to* $\mathcal{C}^2$*, or* $\mathcal{C}^2$ *is* inferior *to* $\mathcal{C}^1$*.*

2. *If* $\{\{j \in N : \hat{\beta}_j = \hat{\beta}_k\} : k \in N\}$ *is a coarser partition than* $\{\{j \in N : \beta_j = \beta_k\} : k \in N\}$*, i.e., each element of the latter collection is a subset of some element of the former collection, we say that the pattern of* $\mathcal{C}^1$ *is* coarser *than that of* $\mathcal{C}^2$*, or the pattern of* $\mathcal{C}^2$ *is* finer *than that of* $\mathcal{C}^1$*. Similarly,* $\mathcal{C}^1$ *and* $\mathcal{C}^2$ *have the same pattern if* $\{\{j \in N : \hat{\beta}_j = \hat{\beta}_k\} : k \in N\} = \{\{j \in N : \beta_j = \beta_k\} : k \in N\}$*.*

We provide the following result by merging two results from [35]. Although they only serve as necessary conditions for a convex cover to be F.I., we will see in Section 2.3 and Section 2.4.1 that the following combination is very powerful.

Theorem 2.2.4. *Let* $\mathcal{C}^1 = \mathcal{C}_{\hat{\boldsymbol{\beta}},\hat{\beta}_0}$ *and* $\mathcal{C}^2 = \mathcal{C}_{\boldsymbol{\beta},\beta_0}$ *be two distinct convex covers on plot* $\mathcal{P}$*. If one of the following is true,*

(i) $\mathcal{C}^1(\bullet) \leq \mathcal{C}^2(\bullet)$;

(ii) $T(\mathcal{C}^1) \supseteq T(\mathcal{C}^2)$*, and the pattern of* $\mathcal{C}^1$ *is* coarser *than that of* $\mathcal{C}^2$;

then $\mathcal{C}^1$ *is superior to* $\mathcal{C}^2$*, i.e.,* $M(\mathcal{C}^1) \supseteq M(\mathcal{C}^2)$*; and moreover,* $\mathcal{C}^2$ *is not F.I.*

Proof. We firstly show that if $M(\mathcal{C}^1) \supseteq M(\mathcal{C}^2)$, then $\mathcal{C}^2$ is not F.I. Suppose otherwise, then $y = \boldsymbol{\beta}^T\boldsymbol{x} + \beta_0$ is a facet, which goes through a set of $(n+1)$ affinely independent extreme points of P, and $y = \hat{\boldsymbol{\beta}}^T\boldsymbol{x} + \hat{\beta}_0$ goes through at least the same set of extreme points of P, so these two hyperplanes are identical, then contradicting the condition that $\mathcal{C}^1$ and $\mathcal{C}^2$ are distinct.

For (i), we know that $0 \leq \mathcal{C}_{\hat{\boldsymbol{\beta}},\hat{\beta}_0}(\bullet) - \mathcal{P}(\bullet) \leq \mathcal{C}_{\boldsymbol{\beta},\beta_0}(\bullet) - \mathcal{P}(\bullet)$. Now consider a level-$k$ vertex $\boldsymbol{a} \in M(\mathcal{C}^2)$, i.e., the level-$k$ extreme point $(\boldsymbol{a}, y^k)$ of P is on $y = \boldsymbol{\beta}^T\boldsymbol{x} + \beta_0$ for some $k \in [n]$. If $\boldsymbol{a} = \boldsymbol{E}_k$, we can imply that $\mathcal{C}_{\hat{\boldsymbol{\beta}},\hat{\beta}_0}(k) = \mathcal{C}_{\boldsymbol{\beta},\beta_0}(k) = \mathcal{P}(k)$, then $(\boldsymbol{E}_k, y^k)$ is obviously on $y = \hat{\boldsymbol{\beta}}^T\boldsymbol{x} + \hat{\beta}_0$. Otherwise, we know $1 \leq k \leq n-1$, and $\boldsymbol{a}$ differs from $\boldsymbol{E}_k$. Then

$$0 = \boldsymbol{\beta}^T\boldsymbol{a} + \beta_0 - y^k \geq \boldsymbol{\beta}^T\boldsymbol{E}_k + \beta_0 - y^k = \mathcal{C}_{\boldsymbol{\beta},\beta_0}(k) - \mathcal{P}(k) \geq \mathcal{C}_{\hat{\boldsymbol{\beta}},\hat{\beta}_0}(k) - \mathcal{P}(k) \geq 0,$$

so that each term equals to 0, moreover, $\boldsymbol{\beta}^T\boldsymbol{a} = \boldsymbol{\beta}^T\boldsymbol{E}_k$. Now let i and j be the first and last entry that $\boldsymbol{a}$ differs from $\boldsymbol{E}_k$, so clearly $1 \le i \le k < j \le n$ and $a_i = l$, $a_j = u$. Notice that $\beta_1 \le \ldots \le \beta_n$, $\boldsymbol{\beta}^T\boldsymbol{a} \ge \boldsymbol{\beta}^T(\boldsymbol{a}$ interchanging i-th and j-th entry$) \ge \boldsymbol{\beta}^T\boldsymbol{E}_k$, the two equalities holds if and only if $\beta_i = \ldots = \beta_j$. This means that $\mathcal{C}_{\boldsymbol{\beta},\beta_0}$ is linear on $\{i-1,\ldots,j\}$, and $\mathcal{C}_{\boldsymbol{\beta},\beta_0}(k) = \mathcal{P}(k)$ where $i \le k \le j-1$. Since $\mathcal{C}_{\hat{\boldsymbol{\beta}},\hat{\beta}_0}(\bullet) \le \mathcal{C}_{\boldsymbol{\beta},\beta_0}(\bullet)$ and $\mathcal{C}_{\hat{\boldsymbol{\beta}},\hat{\beta}_0}(k) = \mathcal{P}(k)$ as well, we must have that $\mathcal{C}_{\hat{\boldsymbol{\beta}},\hat{\beta}_0}(\bullet) = \mathcal{C}_{\boldsymbol{\beta},\beta_0}(\bullet)$ on $\{i-1,\ldots,j\}$, thus $\hat{\beta}_i = \ldots = \hat{\beta}_j$, and hence

$$\hat{\boldsymbol{\beta}}^T\boldsymbol{a} + \hat{\beta}_0 - y^k = \hat{\boldsymbol{\beta}}^T\boldsymbol{E}_k + \hat{\beta}_0 - y^k = \mathcal{C}_{\hat{\boldsymbol{\beta}},\hat{\beta}_0}(k) - \mathcal{P}(k) = 0,$$

which means that $(\boldsymbol{a}, y^k)$ of P is on $y = \hat{\boldsymbol{\beta}}^T\boldsymbol{x} + \hat{\beta}_0$, i.e., $\boldsymbol{a} \in M(\mathcal{C}^1)$. This means $\mathcal{C}^1$ is superior to $\mathcal{C}^2$.

For (ii), consider a level-k vertex $\boldsymbol{a} \in M(\mathcal{C}^2)$; this means that $k \in T(\mathcal{C}^2) \subseteq T(\mathcal{C}^1)$. As $\boldsymbol{a}$ is produced from $\boldsymbol{E}_k$ via permutation, this means that the pattern of $\mathcal{C}^2$ permits $\boldsymbol{a}$ to be produced via permutation; since $\mathcal{C}^1$ has a coarser pattern, it permits $\boldsymbol{a}$ to be produced via permutation, i.e., $\boldsymbol{a} \in M(\mathcal{C}^1)$. This means $\mathcal{C}^1$ is superior to $\mathcal{C}^2$. □

Theorem 2.2.4(i) indicates that an F.I. convex cover should be as low as possible; moreover, Theorem 2.2.4(ii) suggests that the pattern should be as coarse as possible. From the above argument, we can summarize the following result:

Proposition 2.2.5. *$\mathcal{C}$ is F.I. only if there does not exist a convex cover $\mathcal{C}^1$ which is distinct from and superior to $\mathcal{C}$, i.e., $\mathcal{C}(\bullet) \ne \mathcal{C}^1(\bullet)$ and $M(\mathcal{C}) \subseteq M(\mathcal{C}^1)$.*

The reverse of this proposition is conjectured to be true, yet hardly useful, as this kind of non-existence condition is not easy to check.

Example 1. *Figure 2.1 represents the plot for the SMP $-x_1x_2x_3x_4$, and three of its convex covers. Their corresponding inequalities are*

$$\begin{aligned}
C1: &\quad 1 &&\ge y\\
C2: &\quad -x_1 - x_2 + x_3 + x_4 + 3 &&\ge y\\
C3: &\quad -\tfrac{1}{2}x_1 - \tfrac{1}{2}x_2 + \tfrac{1}{2}x_3 + \tfrac{1}{2}x_4 + 2 &&\ge y,
\end{aligned}$$

none of which dominate another in the traditional sense. The patterns for the three covers are listed in Table 2.1. To identify potential F.I. covers, we compare among the three:

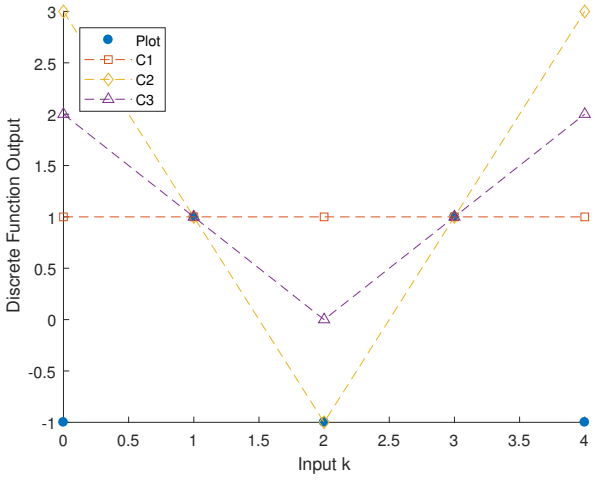

Figure 2.1: Three convex covers for $-x_1x_2x_3x_4$ on $[-1,1]^4$

Convex Cover	Pattern
C1	{1,2,3,4}
C2	{1,2}{3,4}
C3	{1,2}{3,4}

Table 2.1: Pattern of three convex covers

- *C2 and C3 have the same pattern, but C2 goes through more plot points, so C2 is superior to C3 by Theorem* 2.2.4(ii)*;*
- *C1 and C3 go through the same set of plot points, but C1 has a strictly coarser pattern, so C1 is superior to C3 by Theorem* 2.2.4(ii) *as well.*

To argue whether a potential convex cover is indeed F.I.:

- *C1 is indeed F.I., because it is a linear cover that touches the polytope at two level of vertices, by Proposition 2.2.3;*
- *C2 is indeed F.I., because its restrictions on {0,1,2} and {2,3,4} are both linear, hence both F.I. by Proposition 2.2.3; it is F.I. by the forthcoming Theorem* 2.2.6(i).

The forthcoming result is the necessary and sufficient condition for a nonlinear convex cover to be F.I., and serves as the counter part of Proposition 2.2.3. We know that when $\mathcal{P}$, the plot, is not strictly convex, the pattern of item 3 in Definition 6 for an F.I. convex cover cannot be of all singletons. Then it is natural to ask, for a convex cover, how does the pattern lead to it being F.I. Before we prove our main result, we have the following definitions.

Definition 8. *A convex cover $\mathcal{C}$ on plot $\mathcal{P}$, both defined on $[n]$, is* weak F.I.*, or just* weak*, if is not F.I. until either* 0 *or* n *is included in* $T(\mathcal{C})$*. More specifically,*

1. $\mathcal{C}$ *is* L-weak *if it is not F.I. until* 0 *is included in* $T(\mathcal{C})$;

2. $\mathcal{C}$ *is* R-weak *if it is not F.I. until* n *is included in* $T(\mathcal{C})$.

Theorem 2.2.6. *Let* $\mathcal{C} = \mathcal{C}_{\boldsymbol{\beta},\beta_0}$ *be a convex cover on plot* $\mathcal{P}$*, which are both defined on* $[n]$*. Let* $A = \{0, \ldots, i\}, B = \{i, \ldots, n\} \subseteq [n]$ *be two sets of successive integers for some* $1 \leq i \leq n-1$*, and let* $\beta_i < \beta_{i+1}$*. Then* $\mathcal{C}$ *is F.I. on* $\mathcal{P}$ *if and only if one of the following is true:*

(i) $\mathcal{C}|_A$ *is F.I. on* $\mathcal{P}|_A$ *and* $\mathcal{C}|_B$ *is F.I. on* $\mathcal{P}|_B$;

(ii) $\mathcal{C}|_A$ *is F.I. on* $\mathcal{P}|_A$ *and* $\mathcal{C}|_B$ *is L-weak on* $\mathcal{P}|_B$;

(iii) $\mathcal{C}|_A$ *is R-weak on* $\mathcal{P}|_A$ *and* $\mathcal{C}|_B$ *is F.I. on* $\mathcal{P}|_B$.

Proof. ("If") Assume (i). Let $\boldsymbol{v}^1, \ldots, \boldsymbol{v}^{i+1}$ be the $(i+1)$ affinely independent vertices in $M(\mathcal{C}|_A) \subset \mathbb{R}^i$, so the following $(i+1) \times (i+1)$ matrix

$$\begin{bmatrix} 1 & \ldots & 1 \\ \boldsymbol{v}^1 & \ldots & \boldsymbol{v}^{i+1} \end{bmatrix} \tag{2.6}$$

is invertible. The corresponding vertices of X in $\mathbb{R}^n$, each constructed by adding $(n-i)$ entries of l to the tail of their ancestor ($\boldsymbol{v}^j$ for each $1 \leq j \leq i+1$), preserve to be in $M(\mathcal{C})$.

Let $\boldsymbol{v}^{i+2}, \ldots, \boldsymbol{v}^{n+1}$ be a collection of $(n-i)$ affinely independent vertices in $M(\mathcal{C}|_B) \subset \mathbb{R}^{n-i}$, which is still affinely independent when $l\mathbb{1} \in \mathbb{R}^{n-i}$ — the vector with all l — is included in $M(\mathcal{C}|_B)$, so that the following $(n-i) \times (n-i)$ matrix

$$\begin{bmatrix} \boldsymbol{v}^{i+2} - l\mathbb{1} & \ldots & \boldsymbol{v}^{n+1} - l\mathbb{1} \end{bmatrix} \tag{2.7}$$

is invertible. This can be done because a collection of $(n-i+1)$ affinely independent vertices in $M(\mathcal{C}|_B)$ will form an $(n-i)$-simplex in $\mathbb{R}^{n-i}$, and $l\mathbb{1}$ cannot be on all $(n-i+1)$ of its $(n-i-1)$-faces (apply Cramer's Rule to show by contradiction), so that one can find $(n-i)$ of its vertices, the $(n-i-1)$-face going through said vertices does not go through $l\mathbb{1}$. Again, the corresponding vertices of X in $\mathbb{R}^n$, each constructed by adding i entries of u to the head of their ancestor ($\boldsymbol{v}^j$ for each $i+2 \leq j \leq n$), preserve to be in $M(\mathcal{C})$. We put these $(n+1)$ vertices of interest into the

$(n+1)\times(n+1)$ matrix below; via simple row operations we produce a consequent matrice,

$$\left[\begin{array}{ccc|ccc} 1 & \dots & 1 & 1 & \dots & 1 \\ \boldsymbol{v}^1 & \dots & \boldsymbol{v}^{i+1} & u\mathbb{1} & \dots & u\mathbb{1} \\ l\mathbb{1} & \dots & l\mathbb{1} & \boldsymbol{v}^{i+2} & \dots & \boldsymbol{v}^{n+1} \end{array}\right] \to \left[\begin{array}{ccc|ccc} 1 & \dots & 1 & 1 & \dots & 1 \\ \boldsymbol{v}^1 & \dots & \boldsymbol{v}^{i+1} & u\mathbb{1} & \dots & u\mathbb{1} \\ 0 & \dots & 0 & \boldsymbol{v}^{i+2}-l\mathbb{1} & \dots & \boldsymbol{v}^{n+1}-l\mathbb{1} \end{array}\right] \tag{2.8}$$

which is trivially invertible, where $u\mathbb{1}\in\mathbb{R}^i$ is the vector with entries all u.

(ii) and (iii) are two parallel conditions and the proofs are very similar; here we only show for (ii). We still get $\boldsymbol{v}^1,\dots,\boldsymbol{v}^{i+1}$ to be the $(i+1)$ affinely independent vertices in $M(\mathcal{C}|_A)\subset\mathbb{R}^i$, and hence the invertible matrix (2.6). Now that $\mathcal{C}|_B$ is L-weak, $\mathcal{C}|_B$ will be F.I. on $\mathcal{P}|_B$ only when $l\mathbb{1}\in\mathbb{R}^{n-i}$ — the vector with all l — is included in $M(\mathcal{C}|_B)$. Therefore, we can find $\boldsymbol{v}^{i+2},\dots,\boldsymbol{v}^{n+1}$ to be a collection of $(n-i)$ affinely independent vertices in $M(\mathcal{C}|_B)\subset\mathbb{R}^{n-i}$, which is still affinely independent when $l\mathbb{1}$ is included, and hence the invertible matrix (2.7). Then the same matrix invertibility argument follows as in (2.8).

("Only if") From $\beta_i<\beta_{i+1}$, each level-k element in $M(\mathcal{C})$ only relates to an element in $M(\mathcal{C}|_A)$ when $k\in A$ by removing $(n-i)$ l's from its tail, or an element in $M(\mathcal{C}|_B)$ when $k\in B$ by removing i u's from its head. With $\mathcal{C}$ being F.I., it is not possible for both $\mathcal{C}|_A$ and $\mathcal{C}|_B$ to be non-F.I.; otherwise, there are at most i affinely independent points in $M(\mathcal{C}|_A)$ — from the levels in A for $M(\mathcal{C})$, and there are at most $(n-i)$ affinely independent points in $M(\mathcal{C}|_B)$ — from the levels in B for $M(\mathcal{C})$. Hence at most $i+(n-i)=n$ affinely independent points for $M(\mathcal{C})$, contradicting the assumption that $\mathcal{C}$ is F.I.

Now assume that $\mathcal{C}|_A$ is F.I., but $\mathcal{C}|_B$ is not. For a collection of $(n+1)$ affinely independent points in $M(\mathcal{C})$, it must be the case that $(i+1)$ are from levels in A, $(n-i)$ are from the levels in $B\setminus\{i\}$. The $(n+1)\times(n+1)$ matrix transformation (2.8) indicates that (2.7) is invertible, i.e., the $(n-i)$ vertices in $M(\mathcal{C}|_B)$ that relate to the levels in $B\setminus\{i\}$, together with $l\mathbb{1}\in\mathbb{R}^{n-i}$, form a collection of $(n-i+1)$ affinely independent points. This means that $\mathcal{C}|_B$ is L-weak. □

2.3 Convex Hull Generation for Special SMPs

In this section, we try our geometric treatment on the SMPs from [35]. We translate the derivation of all facets for symmetric polytopes with different characteristics, using the convex cover diagram in 2D, with the help of the conditions we referred to and developed. We will see that, when

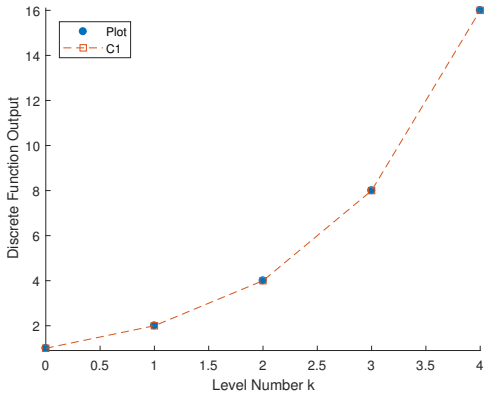

Figure 2.2: F.I. Convex Cover for $\prod_{j=1}^{4} x_j$ on $[1,2]^4$

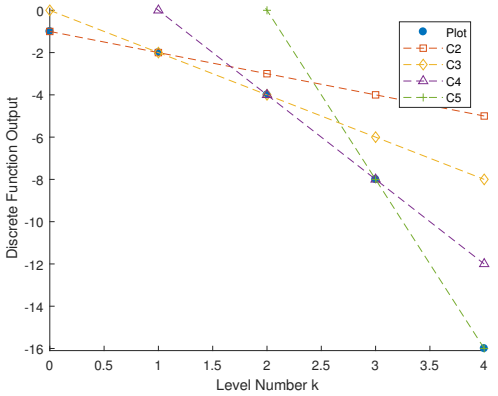

Figure 2.3: F.I. Convex Covers for $-\prod_{j=1}^{4} x_j$ on $[1,2]^4$

plot $\mathcal{P}$ for the symmetric polytope P is convex or concave, i.e., $\mathcal{P}$ has nondecreasing or nonincreasing first forward differences, all the facets of P can be easily derived. Then, we turn to the symmetric polytope generated by the multilinear monomial $m(\boldsymbol{x}) = \prod_{j=1}^{n} x_j$ on $[-1,1]^n$. We simplify the derivation of all the facets for the case on $[-1,1]^n$, and identify a general kind of SMPs that this simplification could handle.

Before getting into details, we need to notice that a plot $\mathcal{P}$ being concave is as good as if it were convex, because it is equivalent as saying that $-\mathcal{P}$, the plot corresponding to the y-value-flipped polytope $-P$, is convex.

2.3.1 Symmetric Polytope with Convex/Concave Plot

Figure 2.2 and Figure 2.3 depict the plot $\mathcal{P}$ for the convex hull of $m(\boldsymbol{x}) = \prod_{j=1}^{n} x_j$ on $X = [1,u]^n$, and all the F.I. convex covers for $\pm\mathcal{P}$. Although the plot is strictly convex for this particular case, according to [35], the strictness of convexity does not affect the identification of F.I. convex covers. In general, the plot of an SMP f is convex if and only if f is supermodular when restricted to the vertices of X; we translate the convex hull generation of Theorem 4.1 in [35] graphically as follows:

Suppose we are provided with a convex plot $\mathcal{P}$, and need to construct the facets for the corresponding polytope P. For the over-estimator side, we can say that $\mathcal{C} = \mathcal{P}$ is a natural convex

cover. By (2.5), its associated facet is

$$\boldsymbol{\beta}^T\boldsymbol{x}+\beta_0=\left(\frac{\mathcal{P}(1)-\mathcal{P}(0)}{u-l},\ldots,\frac{\mathcal{P}(n)-\mathcal{P}(n-1)}{u-l}\right)\boldsymbol{x}+\frac{u\mathcal{P}(0)}{u-l}-\frac{l\mathcal{P}(n)}{u-l}\geq y,$$

and $\{\boldsymbol{E}_k : k\in[n]\}\subseteq M(\mathcal{C})$, the former being facet defining. Then its symmetric, possibly $(n!)$, copies are also facet defining.

The above inequalities are the only facet defining inequalities on the over-estimator side. This is because any convex cover $\mathcal{C}$ other than $\mathcal{P}$ is inferior to $\mathcal{P}$, hence by Theorem 2.2.4, any other possible $\mathcal{C}$ fails to induce a facet.

For the under-estimator side, on the plot $-\mathcal{P}$, we find all the convex covers $\mathcal{C}$ that induce a facet. We notice that the plot $-\mathcal{P}$ here is concave. Consider any nonlinear convex cover $\mathcal{C}$ of $-\mathcal{P}$; they can only touch at a set of successive points, at which points both the nonlinear $\mathcal{C}$ and $-\mathcal{P}$ are linear. Then a linear convex cover through the same collection of points is superior to the nonlinear convex cover $\mathcal{C}$. To identify all linear convex covers that go through a collection of successive points on $-\mathcal{P}$, one can either identify all the collections of successive points at which $-\mathcal{P}$ is linear, OR simply generate all linear convex covers that go through a pair of adjacent points of $-\mathcal{P}$.

Focus on a convex cover that touches the plot at k and $(k+1)$ for some $k\in[n-1]$. The linear convex cover through the two points is superior to any other nonlinear ones. By (2.4) and (2.5),

$$\boldsymbol{\beta}^T\boldsymbol{x}+\beta_0=\frac{-\mathcal{P}(k+1)+\mathcal{P}(k)}{u-l}\left(\sum_{j=1}^{n}x_j-ku-(n-k)l\right)-\mathcal{P}(k)\geq -y,$$

or

$$\boldsymbol{\beta}^T\boldsymbol{x}+\beta_0=\frac{-\mathcal{P}(k+1)+\mathcal{P}(k)}{u-l}\left(\sum_{j=1}^{n}x_j-(k+1)u-(n-k-1)l\right)-\mathcal{P}(k+1)\geq -y.$$

By Proposition 2.2.3, it is facet defining. Thus, any facet defining inequality is obtained for the convex envelope. Translating back from the plot $\mathcal{P}$, we have the following result:

Theorem 2.3.1. *Given any SMP $m(\boldsymbol{x})$ which is supermodular over the extreme points of $X=[l,u]^n$, there exist exactly $(n+1)$ facet-defining inequalities with $\beta_1\leq\ldots\leq\beta_n$ for* conv(G), *and these inequalities are*

$$\frac{um(\boldsymbol{E}_0)-lm(\boldsymbol{E}_n)}{u-l}+\sum_{j=1}^{n}\left(\frac{m(\boldsymbol{E}_j)-m(\boldsymbol{E}_{j-1})}{u-l}\right)x_j-y\geq 0, \tag{2.9}$$

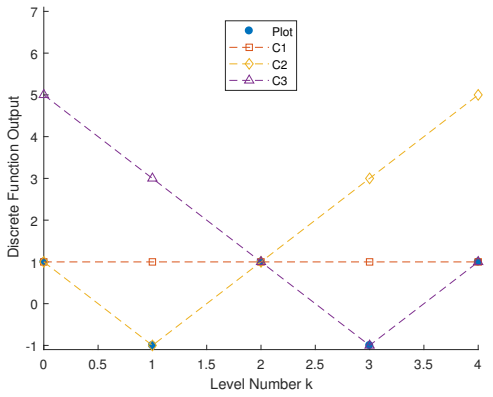

Figure 2.4: F.I. Convex Covers for $\prod_{j=1}^{4} x_j$ on $[-1,1]^4$

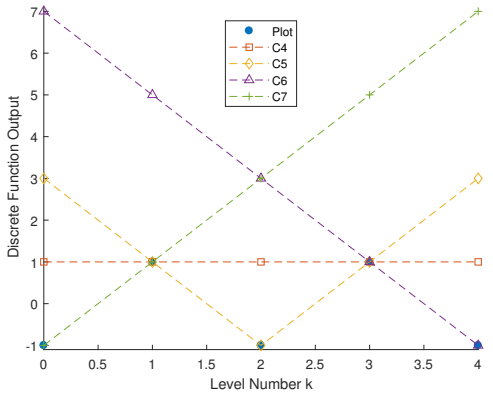

Figure 2.5: F.I. Convex Covers for $-\prod_{j=1}^{4} x_j$ on $[-1,1]^4$

and

$$-m(\boldsymbol{E}_k) - \left(\frac{m(\boldsymbol{E}_k) - m(\boldsymbol{E}_{k-1})}{u-l}\right)\left(\sum_{j=1}^{n} x_j - [ku + (n-k)l]\right) + y \geq 0 \;\forall\, k \in N. \tag{2.10}$$

Though Theorem 2.3.1 is for SMPs over X, as the multilinearity is not used beyond the consequence that conv(G) is a symmetric polytope of (2.2), it reaffirms the convex hull results from [33].

2.3.2 Convex Hull for Monomial on $X = [-1,1]^n$

Figure 2.4 and Figure 2.5 depict the plot $\mathcal{P}$ for the convex hull of the monomial on $X = [-1,1]^n$, and all the F.I. convex covers for $\pm\mathcal{P}$. In fact, regardless of the parity of n, every plot point has a unit absolute height, but with alternating signs as k increases. We simplify the convex hull generation of Section 4.2 in [35] graphically as follows:

To construct the convex cover diagram on $\mathcal{P}(\bullet) = \beta' m(\boldsymbol{E}_\bullet)$ where $\beta' \in \{\pm 1\}$, in order to get the concave envelope inequalities, we need to find all convex covers $\mathcal{C}(\bullet)$ that induce a facet. Another observation is that the convex hull of this plot is either a trapezoid or a parallelogram, hence it is not hard to identify linear F.I. convex covers. We can easily identify the following three kinds of convex covers:

- $\mathcal{C}(\bullet) = 1$ from horizontal edges of the trapezoid or parallelogram. The other horizontal edge

from the flipped plot gives one more F.I. convex cover.

- Two more non-horizontal edges of the trapezoid or parallelogram, flipped plot considered.
- V-shaped convex cover, which consists of exactly two linear portions, and goes through three adjacent points of the heights $1, -1, 1$ respectively. We call it a *Valley*.

On the plot $\mathcal{P}(\bullet) = \beta' m(\boldsymbol{E}_\bullet)$, consider all k such that $\mathcal{P}(k-1) = \mathcal{P}(k+1) = 1$, i.e., $\mathcal{P}(k) = -1$ and $k \notin \{0, n\}$. If $\mathcal{C}(k) \geq 1$ for all such k, then $\mathcal{C}(\bullet) \geq 1$ on $\{1, \ldots, n-1\}$, and even possibly on 0 or n depending upon $\mathcal{P}(\bullet)$; then we must obtain the former two kinds of convex covers, because under the assumption, the plot has become concave. Otherwise, there exists some k such that $\mathcal{P}(k-1) = \mathcal{P}(k+1) = 1$, but $\mathcal{C}(k) < 1$. Then, the V-shaped convex cover $\mathcal{C}^1$, which consists of exactly two linear portions and goes through $(k-1, 1)$ $(k, \mathcal{C}(k))$ and $(k+1, 1)$, is superior to any other convex cover that goes through $(k, \mathcal{C}(k))$; but $\mathcal{C}^1$ is inferior to the Valley through $(k, -1)$ if $\mathcal{C}(k) > -1$; thus the Valley through $(k, -1)$ is potentially F.I. This same argument is in Example 1, shown with Figure 2.1.

A valley is indeed F.I. This fact can be proved by Proposition 2.2.3 and Theorem 2.2.6(i).

An explicit formula for a facet is still recovered by (2.4) and (2.5); notice that the first kind of facet is just $y \in [-1, 1]$, as $\boldsymbol{\beta} = \mathbf{0}$. To illustrate the identification of the other two kinds of facets, consider n even as depicted in Figure 2.4 and Figure 2.5: the second kind of facet only appears in the plot $-\mathcal{P}$; since the slopes for the first two kinds of facets all have rise of ± 2 and run of 2, thus $\beta_j \in \{\pm 1\} \forall j \in N$; the valleys for $\mathcal{P}$, as in Figure 2.4, have an odd number of negative β_j, $j \in N$; the facets for $-\mathcal{P}$, as in Figure 2.5, have an even number of negative β_j, $j \in N$; finally, to apply (2.5) and obtain $\beta_0 = n - 1$, notice that if there are k negative slopes, then the geometry within the diagram indicates that $\mathcal{C}(0) = 2k - 1$ and $\mathcal{C}(n) = 2n - 2k - 1$. Allowing all permutations on $\boldsymbol{\beta}$, we have the following:

$$y \leq -\sum_{j \in T} x_j + \sum_{j \in N \setminus T} x_j + (n-1) \ \forall T \subseteq N \text{ s.t. } |T| \text{ is odd},$$

$$-y \leq -\sum_{j \in S} x_j + \sum_{j \in N \setminus S} x_j + (n-1) \ \forall S \subseteq N \text{ s.t. } |S| \text{ is even}.$$

For an odd n it is analog. Although the plot relies on the parity of n, the collection of all facets can be stated in a concise fashion as the following (2.15), as appeared in [1] and further in [35].

Theorem 2.3.2. *Given any* $m(\boldsymbol{x}) = c_n \prod_{j=1}^{n} x_j$ *having* $X = [-1,1]^n$, *there exist exactly* $(n+3)$ *facet-defining inequalities with* $\beta_1 \leq \ldots \leq \beta_n$ *for* $\mathrm{conv}(G)$, *and these inequalities are as follows:*

$$|c_n| + \beta' y \geq 0 \tag{2.11}$$

for $\beta' = \pm 1$;

$$\beta' \left[-m(\boldsymbol{E}_0) - \left(\frac{m(\boldsymbol{E}_1) - m(\boldsymbol{E}_0)}{2} \right) \left(n + \sum_{j=1}^{n} x_j \right) + y \right] \geq 0, \tag{2.12}$$

where $\beta' = -1$ *if* $m(\boldsymbol{E}_1) > m(\boldsymbol{E}_0)$ *and* $\beta' = 1$ *if* $m(\boldsymbol{E}_1) < m(\boldsymbol{E}_0)$;

$$\beta' \left[-m(\boldsymbol{E}_n) + \left(\frac{m(\boldsymbol{E}_n) - m(\boldsymbol{E}_{n-1})}{2} \right) \left(n - \sum_{j=1}^{n} x_j \right) + y \right] \geq 0, \tag{2.13}$$

where $\beta' = -1$ *if* $m(\boldsymbol{E}_{n-1}) > m(\boldsymbol{E}_n)$ *and* $\beta' = 1$ *if* $m(\boldsymbol{E}_{n-1}) < m(\boldsymbol{E}_n)$; *and*

$$\beta' \left[-m(\boldsymbol{E}_{r+1}) + \left(\frac{m(\boldsymbol{E}_{r+1}) - m(\boldsymbol{E}_r)}{2} \right) \left(-\sum_{j=1}^{r+1} x_j + \sum_{j=r+2}^{n} x_j + n \right) + y \right] \geq 0 \; \forall \, r \in [n-2], \tag{2.14}$$

where $\beta' = -1$ *if* $m(\boldsymbol{E}_{r+1}) < m(\boldsymbol{E}_r)$ *and* $\beta' = 1$ *if* $m(\boldsymbol{E}_{r+1}) > m(\boldsymbol{E}_r)$.

When $c_n = 1$, *we combine and write more succinctly* (2.11), (2.12), (2.13) *and all permutations of* (2.14) *by letting the variable* x_{n+1} *denote* y, *and by letting* $N' = \{1, \ldots, n+1\}$. *These inequalities become*

$$\begin{aligned} -1 \leq & x_j \leq 1 \quad \forall \, j \in N', \\ \sum_{j \in N' \setminus J} x_j - \sum_{j \in J} x_j \leq & (n-1) \quad \forall \, J \subseteq N', |J| \text{ odd}. \end{aligned} \tag{2.15}$$

Upon appealing to the geometry within the convex cover diagram, we notice that the plot of this monomial is special in the sense that it not only is alternating, but also has its plot points lying exactly on the convex hull of the plot $\mathcal{P}$. The exact same argument from above carries over to a more general kind of symmetric polytope P (SMP) as follows:

Proposition 2.3.3. *The symmetric polytope* P *has only two kinds of core facets, whose convex cover is either linear or a valley, if*

$$\{k \in \{1, \ldots, n-1\} : (k, \mathcal{P}(k)) \text{ is on the concave envelope of } \mathcal{P}\}$$

and

$$\{k \in \{1, \ldots, n-1\} : (k, -\mathcal{P}(k)) \text{ is on the concave envelope of } -\mathcal{P}\}$$

partitions $\{1, \ldots, n-1\}$ *by parity.*

2.4 Facet Generation Algorithm and Example

Firstly, we will provide a rather basic procedure that one can implement by hand, mainly in light of Proposition 2.2.3 and Theorem 2.2.4 from [35], and illustrate this procedure through an analytical convex hull generation for a non-supermodular SMP.

Algorithm 1 Extend One-point-a-time

Require: i, j with $0 \leq i < j \leq n$.
  Store $\mathcal{C}(k) = \frac{\mathcal{P}(j)-\mathcal{P}(i)}{j-i}(k-i) + \mathcal{P}(i)$ for $i \leq k \leq j$. STOP
  if there is some $k \in \{i, \ldots, j\}$ with $\mathcal{C}(k) < \mathcal{P}(k)$ **then**
    STOP No convex cover would return
  else
    while $i \geq 1$ **or** $j \leq n-1$ **do**
      if $i \geq 1$ **then**
        Extend the partial cover $\mathcal{C}$ on $\{i-1, \ldots, j\}$ by storing $\mathcal{C}(i-1) = \max\{\mathcal{P}(i-1), 2\mathcal{C}(i) - \mathcal{C}(i+1)\}$, update $i = i-1$;
      end if
      if $j \leq n-1$ **then**
        Extend the partial cover $\mathcal{C}$ on $\{i, \ldots, j+1\}$ by storing $\mathcal{C}(j+1) = \max\{\mathcal{P}(j+1), 2\mathcal{C}(j) - \mathcal{C}(j-1)\}$, update $j = j+1$;
      end if
    end while
    return $\mathcal{C}$
  end if

Algorithm 1 is a linear-time algorithm for quick facet generation, if the structure of $\mathcal{P}$ is not nice. Especially for i, j such that either $\mathcal{C}(i-1) = 2\mathcal{C}(i) - \mathcal{C}(i+1)$ or $\mathcal{C}(j+1) = 2\mathcal{C}(j) - \mathcal{C}(j-1)$ could take place, an F.I. convex cover is guaranteed. However, it does not generate all the facets for us in general, because while it is spreading over the plot, it only chooses the lowest $\mathcal{C}$ values possible; many kinds of F.I. convex covers would be missed by this algorithm, for example, the F.I. convex covers with "wrenches," as discussed in [34].

2.4.1 Basic By-Hand Procedure

We summarize the three steps for the general procedure of facet generation:

1. Construct the Convex Hull of the plot to obtain all the linear convex covers, i.e., the symmetric facets.

2. Search in the plot for easy-to-identify *convex-components* — a collection of plot points whose piecewise linear interpolation is a convex cover; each one gives a face, not necessarily a facet, but may be a good cut.

3. Identify F.I. convex covers intelligently by refining search criterion: each convex-component should be violated to induce a facet.

This procedure relies mainly on Proposition 2.2.3 and Theorem 2.2.4 from [35] to eliminate non-F.I. convex covers and single out potential F.I. convex covers, and then employs other advances for F.I. certification. Though it may be simple and vague, it still provides us insights beyond what the special SMPs from the previous section could. We will use an example to illustrate that the plot structure is a critical factor on the set of all F.I. convex covers, and that as long as the plot structure is not very complicated, even with n being arbitrary, one can handle the analytical calculations on that plot by hand.

Example 2. *Consider* $m(\boldsymbol{x}) = \sum_{i=1}^{n} \prod_{j \neq i} x_j - c \prod_{j=1}^{n} x_j$ *on the unit hypercube,* $\forall c > 0$. *Notice in this situation,*

$$\mathcal{P}(k) = \begin{cases} n - c & \text{if } k = n, \\ 1 & \text{if } k = n-1, \\ 0 & \text{if } k \leq n-2, \end{cases}$$

the plot is not convex as long as $n - c < 2$, *i.e.,* $c > n - 2$. *Now the structure of* P, *or the number of facets, depends on the choice of* c. *We study the following cases for the concave envelope of* P*:*

(i) $n - c < \frac{1}{n-1} + 1$, *i.e.,* $c > n - 1 - \frac{1}{n-1}$, *see Figure 2.6;*

(ii) $n - c = \frac{1}{j} + 1$, *i.e.,* $c = n - 1 - \frac{1}{j}$ *for some* $j \in \{2, \ldots, n-1\}$, *see Figure 2.8;*

(iii) $n - c \in (\frac{1}{j} + 1, \frac{1}{j-1} + 1)$, *i.e.,* $c \in (n - 1 - \frac{1}{j-1}, n - 1 - \frac{1}{j})$ *for some* $j \in \{2, \ldots, n-1\}$, *see Figure 2.10,*

and the following cases for the concave envelope of $-P$*:*

(iv) $c - n < 0$, *i.e.,* $n > c > n - 2$, *see Figure 2.7;*

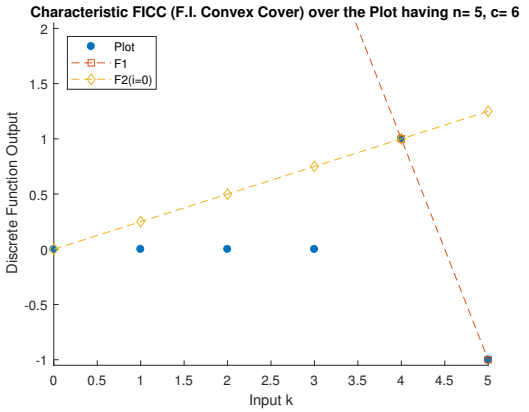

Figure 2.6: Case (i) Plot $\mathcal{P}$

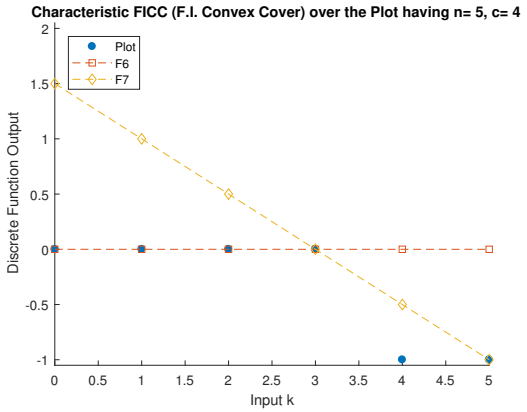

Figure 2.7: Case (iv) Plot $-\mathcal{P}$

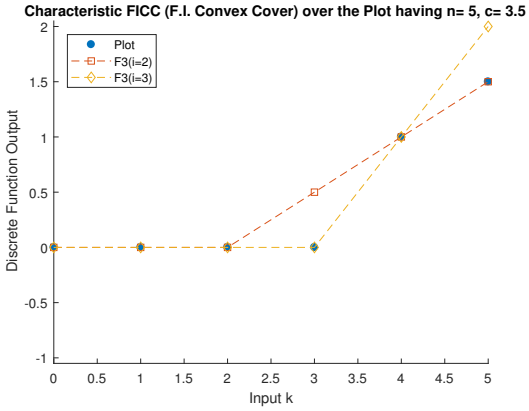

Figure 2.8: Case (ii) Plot $\mathcal{P}$

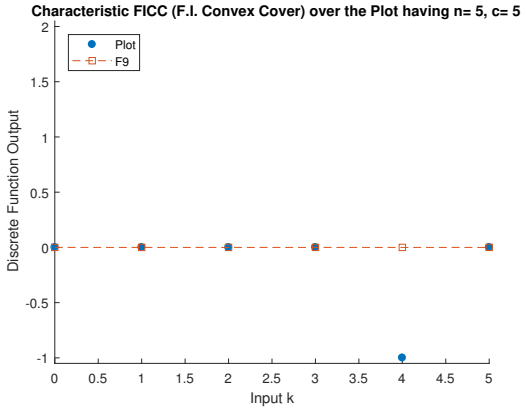

Figure 2.9: Case (v) Plot $-\mathcal{P}$

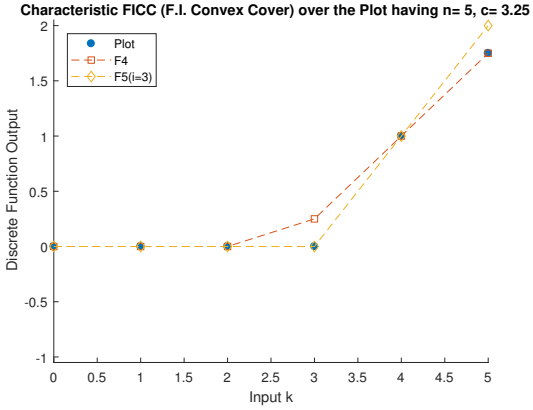

Figure 2.10: Case (iii) Plot $\mathcal{P}$

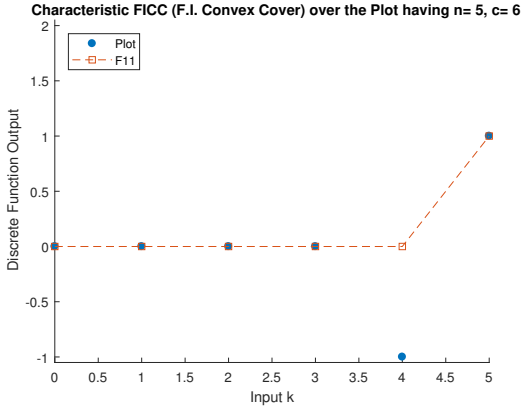

Figure 2.11: Case (vi) Plot $-\mathcal{P}$

(v) $c - n = 0$, *i.e.,* $c = n$, *see Figure 2.9;*

(vi) $c - n > 0$, *i.e.,* $c > n$, *see Figure 2.11.*

Case (i). When $n - c < \frac{1}{n-1} + 1$, *the concave envelope of the plot* $\mathcal{P}$ *gives two linear F.I. covers. We first list the facets:*

$$\begin{aligned} F1: \quad & (n-c-1)(\textstyle\sum_{k=1}^{n} x_k - (n-1)) + 1 \quad \geq y \\ F2: \quad & \qquad\qquad \tfrac{1}{n-i-1} \textstyle\sum_{k=i+1}^{n} x_k \quad \geq y, \forall i \in [n-2]. \end{aligned}$$

To argue the identification of all facets, we start with a convex cover $\mathcal{C}$. *Suppose* i *is the biggest element in* $[n-2] \cap T(\mathcal{C})$; *if such* i *does not exist, then* $T(\mathcal{C}) \subseteq \{n-1, n\}$, *and* $\mathcal{C}$ *is inferior to the linear F.I. cover with the touch set* $\{n-1, n\}$; *if* $i = 0$, *then* $\mathcal{C}(0) = 0$ *and* $\mathcal{C}(n-1) \geq 1$ *implies that* $\mathcal{C}(n) \geq \frac{1}{n-1} + 1$, *which means that* $T(\mathcal{C}) \subseteq \{0, n-1\}$, *and* $\mathcal{C}$ *is inferior to the other linear F.I. cover with the touch set* $\{0, n-1\}$. *Notice now for* $1 \leq i \leq n-2$ *that* $\mathcal{C}$ *is inferior to the convex* $\mathcal{P}$ *on* $[i]$; *and further,* $\mathcal{C}$ *on* $\{i, \ldots, n\}$ *is inferior to the linear cover through* $(i, 0), (n-1, 1)$, *because* $\mathcal{C}(i) = 0, \mathcal{C}(n-1) \geq 1$ *implies that* $\mathcal{C}(n) \geq \frac{1}{n-i-1} + 1$, *which means that* $\{i, \ldots, n\} \cap T(\mathcal{C}) \subseteq \{i, n-1\}$. *Thus the potential F.I. cover must be* 0 *on* $[i]$ *and linear through* $(i, 0), (n-1, 1)$ *on* $\{i, \ldots, n\}$; *the cover is indeed F.I. by Theorem* 2.2.6(i).

Case (ii). When $n - c = \frac{1}{j} + 1$ *for some* $j \in \{2, \ldots, n-1\}$, *we apply what we learned from Case (i): it is easy to identify that a collection of facets remains to be facets:*

$$F3: \quad \tfrac{1}{n-i-1} \textstyle\sum_{k=i+1}^{n} x_k \quad \geq y, \forall i \in \{n-j-1, \ldots, n-2\}.$$

To argue the identification of all facets, we notice that the F.I. cover, which is 0 *on* $[n-j-1]$ *and linear through* $(n-j-1, 0), (n-1, 1)$ *on* $\{n-j-1, \ldots, n\}$, *goes through* $(n, \mathcal{P}(n)) = (n, \frac{1}{j} + 1)$, *so because of Theorem 2.2.4, any other F.I. cover* $\mathcal{C}$ *must be lower on some place in* $\{n-j, \ldots, n-2\}$. *Picking the smallest such* i, *it is not hard to see that* $\mathcal{C}$ *will be inferior if it is not linear on* $\{i, \ldots, n\}$ *through* $(n-1, 1)$; *it will be inferior if it is not* 0 *on* $[i-1]$; *moreover, it will be inferior on* $\{i-1, \ldots, n\}$ *if* $\mathcal{C}(i) \neq 0$. *Eventually it leads to an identified facet. (In Figure 2.8,* $(n-j-1) = 2$.*)*

Case (iii). When $n - c \in (\frac{1}{j} + 1, \frac{1}{j-1} + 1)$, *for some* $j \in \{2, \ldots, n-1\}$, *we apply what we learned from Case (i): it is easy to identify that a collection of facets remain to be facets, with one*

more:

$$\begin{aligned} F4: \quad & x_{n-j} + (n-c-1)(-(j-1)x_{n-j} + \textstyle\sum_{k=n-j+1}^{n} x_k) && \geq y \\ F5: \quad & && \tfrac{1}{n-i-1} \textstyle\sum_{k=i+1}^{n} x_k \geq y, \forall i \in \{n-j, \ldots, n-2\}. \end{aligned}$$

To argue the identification of all facets, we notice that the F.I. cover of F4 has its touch set equal to $[n-j-1] \cup \{n-1, n\}$*, so by Theorem 2.2.4, any other F.I. cover* $\mathcal{C}$ *must be lower on some place in* $\{n-j, \ldots, n-2\}$*. Picking the smallest such* i*, it is not hard to see that* $\mathcal{C}$ *will be inferior if it is not linear on* $\{i, \ldots, n\}$ *through* $(n-1, 1)$*; it will be inferior if it is not* 0 *on* $[i-1]$*; moreover, it will be inferior on* $\{i-1, \ldots, n\}$ *if* $\mathcal{C}(i) \neq 0$*. (In Figure 2.10,* $(n-j-1) = 2$*.)*

Case (iv). When $c - n < 0$*, the concave envelope of the plot* $-\mathcal{P}$ *gives two facets. Notice that* $-\mathcal{P}$ *on* $\{n-2, n-1, n\}$ *has a convex-component, which gives a valley.*

$$\begin{aligned} F6: \quad & 0 && \leq y \\ F7: \quad & \tfrac{n-c}{2}(\textstyle\sum_{k=1}^{n} x_k - (n-2)) && \leq y \\ F8: \quad & \textstyle\sum_{k=1}^{n-1} x_k + (n-c-1)x_n - (n-2) && \leq y. \end{aligned}$$

To argue the identification of all facets, we notice that any nonlinear convex cover is F.I. if $\mathcal{C}(n-1) < \frac{c-n}{2}$*. Such* $\mathcal{C}$ *is inferior if it is not linear on* $[n-1]$ *and through* $(n-2, 0)$*; is inferior if* $\mathcal{C}(n) \neq c-n$*; is inferior if it is not a valley generated on* $\{n-2, n-1, n\}$*.*

Case (v). When $c - n = 0$*, the symmetric facets from Case (iv) coincide. Still,* $-\mathcal{P}$ *on* $\{n-2, n-1, n\}$ *has a convex-component, which gives a valley.*

$$\begin{aligned} F9: \quad & 0 && \leq y \\ F10: \quad & \textstyle\sum_{k=1}^{n-1} x_k - x_n - (n-2) && \leq y. \end{aligned}$$

To argue the identification of all facets, we notice that any nonlinear convex cover is F.I. if $\mathcal{C}(n-1) < 0$*. Such* $\mathcal{C}$ *is inferior if it is not a valley generated on* $\{n-2, n-1, n\}$*.*

Case (vi). When $c - n > 0$*,* $-\mathcal{P}$ *on* $\{n-2, n-1, n\}$ *has a convex-component, which gives a valley; another convex-component is on* $[n] \setminus \{n-1\}$*, which gives another facet:*

$$\begin{aligned} F11: \quad & (n-c)x_n && \leq y \\ F12: \quad & \textstyle\sum_{k=1}^{n-1} x_k + (n-c-1)x_n - (n-2) && \leq y. \end{aligned}$$

To argue the identification of all facets, we notice that any convex cover other than F12 is F.I. if $\mathcal{C}(n-1) < 0$. *Such* $\mathcal{C}$ *is inferior if it is not linear on* $[n-1]$ *and through* $(n-2, 0)$; *is inferior if* $\mathcal{C}(n) \neq c - n$; *is inferior if it is not a valley generated on* $\{n-2, n-1, n\}$.

Summarizing all the cases on a number line $c > n-2$, *we can see how the structure of the convex hull varies as* c *alters.*

- *If* $c \in (n-1-\frac{1}{j-1}, n-1-\frac{1}{j})$ *for some* $j \in \{2, \ldots, n-1\}$, *then it falls under Case (iii) and (iv), therefore the facets are all the symmetric copies of F4, F5, F6, F7, F8;*
- *If* $c = n-1-\frac{1}{j}$ *for some* $j \in \{2, \ldots, n-1\}$, *then it falls under Case (ii) and (iv), therefore the facets are all the symmetric copies of F3, F6, F7, F8;*
- *If* $n > c > n-1-\frac{1}{n-1}$, *then it falls under Case (i) and (iv), therefore the facets are all the symmetric copies of F1, F2, F6, F7, F8;*
- *If* $c = n$, *then it falls under Case (i) and (v), therefore the facets are all the symmetric copies of F1, F2, F9, F10;*
- *If* $c > n$, *then it falls under Case (i) and (vi), therefore the facets are all the symmetric copies of F1, F2, F11, F12.*

The key observation is that the intercept between $k = n$ and any line through two plot points of $\mathcal{P}|_{[n-1]}$ serves as a break point on the number line $c > n-2$. This is no coincidence. For more insights, see [34].

Chapter 3

Exactness of Facet-Defining Inequalities

We quote the following theorem established by Rikun, whose proof is actually one sentence:

Theorem 3.0.1 (Rikun)**.** *The graph of a multilinear* function *over a box has a polyhedral convex hull, whose extreme points relate to vertices of the box.*

Proof. For any $\boldsymbol{x}$ not being a vertex of the box, $(\boldsymbol{x}, g(\boldsymbol{x}))$ on the graph of the multilinear function g is the midpoint of $(\boldsymbol{x}+\epsilon\mathbf{e}_i, g(\boldsymbol{x}+\epsilon\mathbf{e}_i))$ and $(\boldsymbol{x}-\epsilon\mathbf{e}_i, g(\boldsymbol{x}-\epsilon\mathbf{e}_i))$ for some entry i and some $\epsilon > 0$. □

Recall that any valid inequality for the graph of a multilinear polynomial over a box essentially is another nonnegative multilinear polynomial over the entire box. We will see that for the analysis of this chapter, no more information is needed than the previous basic result for multilinear *functions* and some basic knowledge on polynomial algebra that we established at the end of the introduction.

Proposition 3.0.2. *A multilinear function over a box is a multilinear polynomial, i.e., it agrees with some multilinear polynomial on the entire box. If it is also nonnegative, then the following are equivalent:*

(i) *it vanishes at an interior point of the box;*

(ii) *it vanishes at each vertex of the box;*

(iii) *it vanishes on the entire box.*

Proof. Let $g(\boldsymbol{x})$ be a multilinear function on X' of (1.4). It agrees with

$$p(\boldsymbol{x}) = \frac{1}{\prod_{j=1}^{n}(U_j - L_j)} \sum_{J \subseteq N} g(\boldsymbol{E}_J) F(J)$$

defined from (1.7) at every vertex of X'. Then of course $g - p$ is a multilinear function, and vanishes at every vertex of X'. By Theorem 3.0.1, its convex hull is $y = 0$, which implies that $g = p$ on X'.

The above argument also shows (ii)⇒(iii). (iii)⇒(i) is obvious. To show (i)⇒(ii), we utilize the result above, that g can be expressed as $\frac{1}{\prod_{j=1}^{n}(U_j - L_j)} \sum_{J \subseteq N} g(\boldsymbol{E}_J) F(J)$ on X'. Since at an interior point of X', $F(J)$ is positive for all $J \subseteq N$, g must vanish at each vertex of X'. □

As discussed in Section 1.4, the functional values of a multilinear polynomial over the vertices of X' of (1.4) completely determine its nonnegativity over X'. They, as we show next, determine the set of points in X' at which a nonnegative multilinear polynomial vanishes as well. As a matter of fact, a nonnegative multilinear polynomial vanishes on a collection of faces of the box X', which is completely determined by the vertices of X' at which the polynomial vanishes.

Theorem 3.0.3. *Consider multilinear function (polynomial) $g(\boldsymbol{x})$ that is nonnegative over X' of (1.4), and any point $\hat{\boldsymbol{x}} \in X'$. Partition N into N_1, N_2, and N_3 so that $N_1 \equiv \{j : \hat{x}_j = L_j\}$, $N_2 \equiv \{j : \hat{x}_j = U_j\}$, and $N_3 \equiv \{j : L_j < \hat{x}_j < U_j\}$. Then the following are equivalent:*

(i) $g(\hat{\boldsymbol{x}}) = 0$;

(ii) $g(\tilde{\boldsymbol{x}}) = 0$ *for all vertices* $\tilde{\boldsymbol{x}} \in X'$ *having* $\tilde{x}_j = L_j \ \forall\, j \in N_1$ *and* $\tilde{x}_j = U_j \ \forall\, j \in N_2$;

(iii) $g(\tilde{\boldsymbol{x}}) = 0$ *for all* $\tilde{\boldsymbol{x}} \in X'$ *having* $\tilde{x}_j = L_j \ \forall\, j \in N_1$ *and* $\tilde{x}_j = U_j \ \forall\, j \in N_2$.

Proof. Upon restricting $x_j = L_j \ \forall j \in N_1$ and $x_j = U_j \ \forall j \in N_2$, $g(\boldsymbol{x})$ on the box in $\mathbb{R}^{N_3}$ is still a nonnegative multilinear function. This result is then evident from Proposition 3.0.2. □

The above theorem indicates that the nonnegative multilinear polynomial vanishes on the entire faces that are constructed from the vertices at which the polynomial vanishes. We derive the points at which each facet from each convex hull form is exact in the following sections.

3.1 Supermodular SMP

The below theorem identifies, for a supermodular SMP $m(\boldsymbol{x})$, the set of all points in G of (1.1) that satisfies inequality (2.9) with equality.

Theorem 3.1.1. *Given a supermodular $m(\boldsymbol{x})$, the facet-defining inequality (2.9) of Theorem 2.3.1 is satisfied with equality at a point $(\hat{\boldsymbol{x}}, \hat{y}) \in G$ if and only if there exists $p, q \in N$ with $p \leq q$ and*

$$m(\boldsymbol{E}_p) - m(\boldsymbol{E}_{p-1}) = \ldots = m(\boldsymbol{E}_q) - m(\boldsymbol{E}_{q-1}),$$

so that $\hat{\boldsymbol{x}}$ has $\hat{x}_j = u$ for all $j < p$, $l \leq \hat{x}_p, \ldots, \hat{x}_q \leq u$, $\hat{x}_j = l$ for all $j > q$.

Proof. For simplicity of the notation, we denote the scalars of core facet (2.9) as $\boldsymbol{\beta}$ and β_0 following the convention of Chapter 2. By definition of (2.9), this facet is satisfied with equality at each $(\boldsymbol{E}_j, m(\boldsymbol{E}_j))$ for all $j \in [n]$.

The "if" direction is straight forward: if such p, q exists, then by $\beta_p = \ldots = \beta_q$ from the construction of (2.9), it is satisfied with equality for each $(\boldsymbol{a}, m(\boldsymbol{a}))$ with $\boldsymbol{a}$ being a vertex having $a_j = u$ for all $j < p$ and $a_j = l$ for all $j > q$. Thus, by Theorem 3.0.3, as $\hat{\boldsymbol{x}}$ lies on the face $x_j = u$ for all $j < p$ and $x_j = l$ for all $j > q$, the equality is satisfied for $(\hat{\boldsymbol{x}}, \hat{y})$.

For the "only if" direction, we first rearrange $\hat{\boldsymbol{x}}$ into $\bar{\boldsymbol{x}}$ that has non-ascending entries, i.e., $\bar{x}_1 \geq \ldots \geq \bar{x}_n$. Since $\bar{\boldsymbol{x}}$ is the minimizer of the affine expression $\boldsymbol{\beta}^T \bullet + \beta_0$ among all permutations of $\hat{\boldsymbol{x}}$, under the same logic of the proof of Proposition 2.2.1, $(\bar{\boldsymbol{x}}, \hat{y})$ must also satisfy the equality, as

$$\hat{y} = \boldsymbol{\beta}^T \hat{\boldsymbol{x}} + \beta_0 \geq \boldsymbol{\beta}^T \bar{\boldsymbol{x}} + \beta_0 \geq \hat{y}.$$

There are two possibilities for $\bar{\boldsymbol{x}}$:

- if $\bar{\boldsymbol{x}} = \boldsymbol{E}_k$ for some $k \in N$, let $p_0 = q_0 = k$; if $\bar{\boldsymbol{x}} = \boldsymbol{E}_0$, let $p_0 = q_0 = 1$;
- otherwise, let p_0 be the first index such that $\bar{x}_{p_0} < u$, and let q_0 be the last index such that $\bar{x}_{q_0} > l$.

In either case, we have

$$\beta_{p_0} = \ldots = \beta_{q_0}.$$

For the latter case especially, since $\bar{\boldsymbol{x}}$ is an interior point of the face $x_j = u$ for all $j < p_0$ and $x_j = l$

for all $j > q_0$, by Theorem 3.0.3, the equality holds for every vertex of this face, including $\boldsymbol{E}_{p_0}$ and $\boldsymbol{E}_{p_0-1} + (u-l)\boldsymbol{e}_{q_0}$. This essentially leads to

$$\boldsymbol{\beta}^T \boldsymbol{E}_{p_0} + \beta_0 = \boldsymbol{\beta}^T \left(\boldsymbol{E}_{p_0-1} + (u-l)\boldsymbol{e}_{q_0}\right) + \beta_0 = m(\boldsymbol{E}_{p_0}),$$

and hence $\beta_{p_0} = \beta_{q_0}$.

Now, if $\hat{\boldsymbol{x}}$ and $\bar{\boldsymbol{x}}$ only differs among the entries from p_0 to q_0, we let $p = p_0$ and $q = q_0$, then the proof is complete. Otherwise, there exists the smallest $p_1 < p_0$ such that $\hat{x}_{p_1} < u$ *or* there exists the largest $q_1 > q_0$ such that $\hat{x}_{q_1} > l$. Notice that if p_1 does exist, we can identify some entry $\hat{x}_i = u$ where $i \geq p_0$, permute $\hat{x}_{p_1}$ and $\hat{x}_i$ within $\hat{\boldsymbol{x}}$ to obtain a vector $\tilde{\boldsymbol{x}}$, which squeezes into

$$\hat{y} = \boldsymbol{\beta}^T \hat{\boldsymbol{x}} + \beta_0 \geq \boldsymbol{\beta}^T \tilde{\boldsymbol{x}} + \beta_0 \geq \boldsymbol{\beta}^T \bar{\boldsymbol{x}} + \beta_0 \geq \hat{y},$$

because the affine expression $\boldsymbol{\beta}^T \bullet + \beta_0$ is lessened from $\hat{\boldsymbol{x}}$ to $\tilde{\boldsymbol{x}}$ yet the minimizer is $\bar{\boldsymbol{x}}$; and therefore $\beta_{p_1} = \ldots = \beta_{p_0}$. Similarly, if q_1 does exist, we will have $\beta_{q_0} = \ldots = \beta_{q_1}$. Let $p = \min\{p_0, p_1\}$, $q = \max\{q_0, q_1\}$, and combine with $\beta_{p_0} = \ldots = \beta_{q_0}$ obtained previously, we have that $\beta_p = \ldots = \beta_q$. $\square$

Theorem 3.1.2 below identifies, for a supermodular SMP $m(\boldsymbol{x})$, the set of all points in G of (1.1) that satisfies inequality (2.10) with equality for any given $k \in N$ with equality.

Theorem 3.1.2. *Given a supermodular $m(\boldsymbol{x})$ and any $k \in N$, the facet-defining inequality* (2.10) *of Theorem 2.3.1 is satisfied with equality at a point $(\hat{\boldsymbol{x}}, \hat{y}) \in G$ if and only if there exists $p, q \in N$ with $p \leq k \leq q$ and*

$$m(\boldsymbol{E}_p) - m(\boldsymbol{E}_{p-1}) = \ldots = m(\boldsymbol{E}_q) - m(\boldsymbol{E}_{q-1}),$$

so that $\hat{\boldsymbol{x}}$ contains $(p-1)$ entries of value u, $(n-q)$ entries of value l, the remaining $(q-p+1)$ entries in the interval $[l, u]$.

Proof. Notice that the core facet (2.10) is a symmetric hyperplane with respect to $\boldsymbol{x}$, and has a linear convex cover, therefore without loss of generality we only show for $\hat{\boldsymbol{x}}$ having non-ascending entries, i.e., $\hat{x}_1 \geq \ldots \geq \hat{x}_n$. By definition of (2.10), this facet is satisfied with equality at $(\boldsymbol{E}_{k-1}, m(\boldsymbol{E}_{k-1}))$ and $(\boldsymbol{E}_k, m(\boldsymbol{E}_k))$.

The "if" direction is straight forward: if such p, q exists, then by the construction of (2.10), since $m(\boldsymbol{E}_p) - m(\boldsymbol{E}_{p-1}) = \ldots = m(\boldsymbol{E}_q) - m(\boldsymbol{E}_{q-1})$, it is satisfied with equality for each $(\boldsymbol{a}, m(\boldsymbol{a}))$

with $\boldsymbol{a}$ being a vertex having $a_j = u$ for all $j < p$ and $a_j = l$ for all $j > q$. Thus, by Theorem 3.0.3, as $\hat{\boldsymbol{x}}$ lies on the face $x_j = u$ for all $j < p$ and $x_j = l$ for all $j > q$, the equality is satisfied for $(\hat{\boldsymbol{x}}, \hat{y})$.

For the "only if" direction, there are two possibilities for $\hat{\boldsymbol{x}}$:

1. if $\hat{\boldsymbol{x}} = \boldsymbol{E}_i$ for some $i \in [n]$,

   - and further if $i \leq k-1$, let $p = i+1$ and $q = k$, the proof is then complete, as in this case the linear cover touches the concave plot at level i, $(k-1)$, and k, hence every level from i to k;
   - if $i \geq k$, let $p = k$ and $q = i$, the proof is also complete, as in this case the linear cover touches the concave plot at level $(k-1)$, k, and i, hence every level from $(k-1)$ to i;

2. otherwise, let p_0 be first index such that $\hat{x}_{p_0} < u$, and let q_0 be the last index such that $\hat{x}_{q_0} > l$.

In the latter case, we have

$$m(\boldsymbol{E}_{p_0}) - m(\boldsymbol{E}_{p_0-1}) = \ldots = m(\boldsymbol{E}_{q_0}) - m(\boldsymbol{E}_{q_0-1}),$$

because $\hat{\boldsymbol{x}}$ is an interior point of the face $x_j = u$ for all $j < p_0$ and $x_j = l$ for all $j > q_0$, by Theorem 3.0.3, the equality holds for every vertex of this face, including $\boldsymbol{E}_{p_0-1}$ and $\boldsymbol{E}_{q_0}$, which means that the linear cover touches the concave plot at level $(k-1)$, k, $p_0 - 1$ and q_0, hence every level from p to q, where $p \equiv \min\{k-1, p_0-1\}$ and $q \equiv \max\{k, q_0\}$. The proof is therefore complete. □

3.2 Monomial on $X = [-1, 1]^n$

Referring back to the generation of three kinds of facets in Section 2.3.2, we can easily identify the extreme points at which the equality is satisfied, then utilizing Theorem 3.0.3, we can easily identify the points at which the equality is satisfied as follows:

- The horizontal covers of $y \leq 1$ and $y \geq -1$ from horizontal edges of the trapezoid or parallelogram do not touch the plot at adjacent levels. No face will be formed from those vertices, as a face at least relates to two vertices from adjacent levels. Therefore these inequalities are satisfied as equalities only at half of the vertices of X.

- Two other linear covers from non-horizontal edges of the trapezoid or parallelogram touch the plot at levels $\{0, 1\}$ and $\{n-1, n\}$, respectively. Therefore by Theorem 3.0.3, a total of n 1-faces will be formed between the level-0 vertex and the level-1 vertices for one facet, and another n between the level-n vertex and the level-$(n-1)$ vertices for the other facet.

- Each facet from a valley, which consists of exactly two linear portions, and goes through three adjacent points $(k-1, 1)$, $(k, -1)$, and $(k+1, 1)$ for some $k \in \{1, \ldots, n-1\}$, is satisfied with equality at the vertex $\boldsymbol{E}_k$ of level-k, k vertices of level-$(k-1)$, and $(n-k)$ vertices of level-$(k+1)$, which all differ from $\boldsymbol{E}_k$ by at most one entry. Since there is only one level-k vertex, no 2-face will be formed from these $(n+1)$ vertices, and hence a total of n 1-faces will be formed between the level-k vertex and the rest.

Theorem 3.2.1 below summarizes, for $m(\boldsymbol{x}) = c_n \prod_{j=1}^{n} x_j$ having $X = [-1, 1]^n$, the set of all points in G of (1.1) that satisfies an inequality as found in (2.11), (2.12), (2.13) or (2.14) with equality.

Theorem 3.2.1. *Given any $m(\boldsymbol{x}) = c_n \prod_{j=1}^{n} x_j$ having $X = [-1, 1]^n$, the facet-defining inequality*

1. (2.11) *of Theorem 2.3.2 is satisfied with equality at a point $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \in G$ if and only if $\hat{\boldsymbol{x}}$ contains an even number of entries of value -1, the remaining entries of value 1, when $c_n \beta' < 0$, or contains an odd number of entries of value -1, the remaining entries of value 1, when $c_n \beta' > 0$;*

2. (2.12) *of Theorem 2.3.2 is satisfied with equality at a point $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \in G$ if and only if $\hat{\boldsymbol{x}}$ contains $(n-1)$ entries of value -1, the remaining entry in the interval $[-1, 1]$;*

3. (2.13) *of Theorem 2.3.2 is satisfied with equality at a point $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \in G$ if and only if $\hat{\boldsymbol{x}}$ contains $(n-1)$ entries of value 1, the remaining entry in the interval $[-1, 1]$;*

4. (2.14) *of Theorem 2.3.2 with $r \in [n-2]$ is satisfied with equality at a point $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \in G$ if and only if $\hat{\boldsymbol{x}}$ contains r entries of value 1, $(n-r-1)$ entries of value -1, the remaining entry in the interval $[-1, 1]$, or contains $(r+1)$ entries of value 1, $(n-r-2)$ entries of value -1, the remaining entry in the interval $[-1, 1]$.*

Remark 5. *The study of the (extreme) points of X at which the graph intersects the facet produces the region on which the facet is "effective" among the convex hull form: a non-vertical facet is either a component of the concave envelope or the convex envelope, therefore the convex hull of the vertices*

of X at which the graph intersects the facet gives the region on which the facet is "lowest" (among the over-estimators) or "highest" (among the under-estimators).

Within (2.15) *(or Theorem 2.3.2), there are 2^n nonhorizontal facets and 2 horizontal ones. From the above theorem and the analysis, we know that each nonhorizontal facet is effective only on a nice n-simplex of X, with a volume of $\frac{2^n}{n!}$, as the vertices at which the facet intersects the graph differ from some vertex by at most one entry. Notice that the total volume of X is 2^n; there are 2^{n-1} nonhorizontal facets among the concave envelope form, therefore within $\left(1-\frac{2^{n-1}}{n!}\right)$ of the volumne of X, $y \leq 1$ is the only effective facet. Similarly, within $\left(1-\frac{2^{n-1}}{n!}\right)$ of the volumne of X, $y \geq -1$ is the only effective facet among the convex envelope. Therefore, at least within $\left(1-\frac{2^n}{n!}\right)$ of the volumne of X, $y \in [-1,1]$ is the only restriction for the approximation using the convex hull. Both $\frac{2^{n-1}}{n!}$ and $\frac{2^n}{n!}$ vanish when n is sufficiently large.*

3.3 SMP from Example 2

In Example 2, the nonsupermodular SMP is defined on the unit hypercube $X = [0,1]^n$. Similar to the construction from the previous section, for the facets obtained from Example 2, it is easy to utilize Theorem 3.0.3 and obtain the following list:

F1: $\boldsymbol{x}$ must contain $(n-1)$ entries of value 0, the remaining entry in the interval $[0,1]$, because the linear cover only touches the plot at level $(n-1)$ and n;

F2: For the parameter i, as the convex cover touches the plot at all levels in $[i]$ while it stays linear on $[i]$, and then touches levels $(n-1)$ while it stays linear on $\{i,\ldots,n\}$:

(a) if $i = n-2$, $\boldsymbol{x}$ must have all 0's in the last $(n-i)$ entries *or* have the first $(n-2)$ entries 1's, and a 0 among the last two entries;

(b) if $i \leq n-3$, $\boldsymbol{x}$ must have all 0's in the last $(n-i)$ entries *or* be a level-$(n-1)$ vertex having the first i entries 1's;

F3: Same facet as F2 except for the parameter range,

(a) if $i = n-2$, $\boldsymbol{x}$ must have all 0's in the last $(n-i)$ entries *or* have the first $(n-2)$ entries 1's, and a 0 among the last two entries;

(b) if $i \leq n-3$, $\boldsymbol{x}$ must have all 0's in the last $(n-i)$ entries *or* be a level-$(n-1)$ vertex having the first i entries 1's;

F4: $\boldsymbol{x}$ must have all 0's in the last $(j+1)$ entries *or* have $(n-1)$ entries of 1's including all the first $(n-j)$ entries, because the convex cover touches the plot at all levels in $[n-j-1]$ while it stays linear on $[n-j-1]$, and then touches levels $(n-1)$ and n while it stays linear on $\{n-j, \ldots, n\}$;

F5: Same facet as F2 and F3 except for the parameter range,

(a) if $i = n-2$, $\boldsymbol{x}$ must have all 0's in the last $(n-i)$ entries *or* have the first $(n-2)$ entries 1's, and a 0 among the last two entries;

(b) if $i \leq n-3$, $\boldsymbol{x}$ must have all 0's in the last $(n-i)$ entries *or* be a level-$(n-1)$ vertex having the first i entries 1's;

F6: $\boldsymbol{x}$ must have two entries of 0's, because the horizontal linear cover touches the plot at all levels in $[n-2]$;

F7: $\boldsymbol{x}$ must be a level-$(n-2)$ vertex *or* be $\boldsymbol{E}_n = \mathbb{1}$, because the linear cover touches the plot at levels $(n-2)$ and n;

F8: $\boldsymbol{x}$ must differ from $\boldsymbol{E}_{n-1}$ by at most one entry, because the valley only touches the plot at level $(n-2)$, $(n-1)$, and n;

F9: $\boldsymbol{x}$ must have two entries of 0's *or* be $\boldsymbol{E}_n = \mathbb{1}$, because the horizontal linear cover touches the plot at all levels in $[n-2] \cup n$;

F10: $\boldsymbol{x}$ must differ from $\boldsymbol{E}_{n-1}$ by at most one entry, same as F8;

F11: $\boldsymbol{x}$ must have two entries of 0's, one of which in the last entry, *or* be $\boldsymbol{E}_n = \mathbb{1}$, because the convex cover touches the plot at all levels in $[n-2]$ while it stays linear on $[n-1]$, and then touches at level n;

F12: $\boldsymbol{x}$ must differ from $\boldsymbol{E}_{n-1}$ by at most one entry, same as F8 and F10.

Chapter 4

Error Analysis of Monomial Convexifications in Polynomial Optimization

In this section, we consider the approximation of a monomial over (a subset of) a box. The linearization technique that we have discussed at the beginning of Section 2.1, is a very common technique in polynomial optimization problems. Consider $p(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha}} c_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}}$ where the sum is finite, $\boldsymbol{x}^{\boldsymbol{\alpha}} \equiv \prod_{j=1}^{n} x_j^{\alpha_j}$ is a monomial, and every α_j is a positive integer; since it is NP-hard to compute the convex envelope of an arbitrary polynomial, it is more practical, cheaper and easier to simply linearize, or convexify (using the possibly non-polyhedral convex hull) each monomial term $\boldsymbol{x}^{\boldsymbol{\alpha}}$ that appeared in the optimization problem. We have already summarized the known monomial convex hulls in the literature in Section 2.1.

4.0.1 Motivation and Literature Review

To quantify the strength of a relaxation of $p(\boldsymbol{x})$, one is interested in bounding the error produced with respect to the global optimum $z_S^* \equiv \min\{p(\boldsymbol{x}) : \boldsymbol{x} \in S\}$ by optimizing over this relaxation. Error bounds for converging solutions of iterative optimization algorithms have been the subject of study before [24], but since these are not suited for studying relaxation strengths,

different error measures have been proposed. Luedtke et al. [19] studied a relative error measure for the relaxation of a bilinear polynomial $p(\boldsymbol{x})$ on $S = [0,1]^n$ obtained by convexifying each monomial with its McCormick envelopes. They showed that for every $\boldsymbol{x} \in [0,1]^n$, the ratio of the difference between the McCormick over-estimator and under-estimator values at $\boldsymbol{x}$ and the difference between the concave and convex envelope values at $\boldsymbol{x}$ can be bounded by a constant that is solely in terms of the chromatic number of the co-occurrence graph of the bilinear polynomial. Recently, Boland et al. [5] showed that this same ratio cannot be bounded by a constant independent of n. Another, and somewhat natural, way of measuring the error from a relaxation is to bound the absolute gap $z_S^* - \tilde{z}_S$, where $\tilde{z}_S$ is a lower bound on z_S^* due to some convex relaxation of $\{(\boldsymbol{x}, y) \in S \times \mathbb{R} : y = p(\boldsymbol{x})\}$. Such a bound helps determine how close one is to optimality in a global optimization algorithm. Also, there are examples (cf. $\prod_{j=1}^n x_j$ on $[1,r]^n$ in [19, pp. 332]) where the relative error gap of McCormick relaxation goes to ∞, while this can never happen with the absolute gap. The only result that we know of on bounding absolute gaps is due to De Klerk and Laurent [12] who used Bernstein approximation of polynomials for a hierarchy of linear program and semidefinite relaxations. (On the contrary, [13, 14] bound the absolute error from upper bounds on z_S^*.). Finally, we mention that a third error measure is based on comparing the volume of a convex relaxation to the volume of the convex hull. This has been done for McCormick relaxations of a trilinear monomial over a box [31].

In this section, we bound the absolute gap to a true polynomial output value $p(\boldsymbol{x})$, from monomial convexification and thereby add to the small number of explicit error bounds for polynomial optimization in the literature. To bound this gap, we analyze the error in relaxing a monomial with its convex hull. This error analysis not only implies a bound on the absolute gap to the true polynomial value but it also can be used for bounding the error in relaxing any optimization problem with polynomials in both objective and constraints. Our error measure is the maximum absolute deviation between the actual value and the approximate value of the monomial. Thus for any set $S \subseteq \mathbb{R}^n$, we denote the worst-case error of the concave envelope, convex envelope, convex hull with respect to $\boldsymbol{x}^{\boldsymbol{\alpha}}$ on S by $\mu^{\text{cav}}(S)$, $\mu^{\text{vex}}(S)$, $\mu(S)$ respectively:

$$\mu(S) \equiv \max \ |y - \boldsymbol{x}^{\boldsymbol{\alpha}}| \ \text{ s.t. } (x,y) \in \text{conv}\{(\boldsymbol{x}, y) \in S \times \mathbb{R} : y = \boldsymbol{x}^{\boldsymbol{\alpha}}\}, \tag{4.1}$$

$$\mu^{\text{cav}}(S) \equiv \max \ y - \boldsymbol{x}^{\boldsymbol{\alpha}} \text{ s.t. } (x,y) \in \text{conv}\{(\boldsymbol{x}, y) \in S \times \mathbb{R} : y \leq \boldsymbol{x}^{\boldsymbol{\alpha}}\}, \tag{4.2}$$

$$\mu^{\text{vex}}(S) \equiv \max \ \boldsymbol{x}^{\boldsymbol{\alpha}} - y \text{ s.t. } (x,y) \in \text{conv}\{(\boldsymbol{x}, y) \in S \times \mathbb{R} : y \geq \boldsymbol{x}^{\boldsymbol{\alpha}}\}. \tag{4.3}$$

We will mostly be interested in the error for the convex hull of the graph of $\boldsymbol{x}^{\boldsymbol{\alpha}}$ and for the convex and concave envelopes of $\boldsymbol{x}^{\boldsymbol{\alpha}}$. Monomial convexification errors have not been addressed before in the literature. The only result in this regard is the folklore result [2] for a bilinear monomial x_1x_2 on $[L_1, U_1] \times [L_2, U_2]$ stating that the convex hull and envelope errors are attained at $(x_1, x_2) = (\frac{U_1+L_1}{2}, \frac{U_2+L_2}{2})$, which is the midpoint of the two diagonals of the box. Since convex hull and envelopes results for a bilinear polynomial are invariant to affine transformations, it is equivalent to consider x_1x_2 on $[0, 1]^2$ and say that the errors are each $\frac{1}{4}$ and are attained at $(\frac{1}{2}, \frac{1}{2})$. Substituting $n = 2, \alpha_1 = \alpha_2 = 1$ in our error bounds recovers this result.

We obtain strong and explicit upper bounds on $\mu(\bullet)$ for different types of monomials. In the polynomial optimization literature, it is common to assume, upto scaling and translation, that the domain S of the problem is a subset of $[0, 1]^n$. When analyzing a single monomial, this assumption is not without loss of generality since the set of monomials is not closed upto translating and scaling the variables. Hence we divide our analysis into two parts. First, we consider a general monomial $\boldsymbol{x}^{\boldsymbol{\alpha}}$ over a compact convex $S \subseteq [0, 1]^n$, and bound the errors without using explicit analytic forms of the envelopes, which are hard to compute and unknown in closed form for arbitrary S. The concave error is bounded by computing the error from a specific concave over-estimator that is precisely the concave envelope of $\boldsymbol{x}^{\boldsymbol{\alpha}}$ on $[0, 1]^n$. On the convex side, we bound the error for any convex under-estimator given as the pointwise supremum of (at most countably many) linear functions, each of which underestimates $\boldsymbol{x}^{\boldsymbol{\alpha}}$ on S.

In the second part, we limit our attention to a multilinear monomial $m(\boldsymbol{x}) = \prod_{j=1}^{n} x_j$, but the domain S is either a box with constant ratio or a box with symmetric bounds. By a box with constant ratio, we mean any box for which there exists a scalar $r > 1$ such that $U_i/L_i = r$ for all i with $L_i > 0$, and $L_i/U_i = r$ for all i with $L_i < 0$. By a box with symmetric bounds, we mean any box that has $U_i = -L_i$ for all i. Since these boxes are simple scalings of $[1, r]^n$ and $[-1, 1]^n$, respectively, and our error measure $\mu(\bullet)$ scales, we restrict our attention to only $[1, r]^n$ and $[-1, 1]^n$. Contrary to the first part, here we first derive explicit polyhedral characterizations of the envelopes and convex hulls on $[1, r]^n$ and $[-1, 1]^n$ and use them to perform a tight error analysis. The polyhedral representations for the $[1, r]^n$ case follow from the literature and is alternatively derived in Section 2.1, whereas those on $[-1, 1]^n$ are established from Section 2.1.

4.0.2 General monomial

Consider a monomial $\boldsymbol{x}^{\boldsymbol{\alpha}}$ with $\alpha_j \in \mathbb{Z}_{\geq 1}$ for all j. The degree of this monomial is $d \equiv |\boldsymbol{\alpha}| = \sum_{j=1}^{n} \alpha_j$. The following constants will be useful for majority of the section:

$$\mathcal{C}_d^1 \equiv \left(1 - \frac{1}{d}\right) d^{\frac{1}{1-d}}, \qquad \mathcal{C}_d^2 \equiv \left(1 - \frac{1}{d}\right)^d.$$

Theorem 4.0.1. *For the monomial $m(x) = \boldsymbol{x}^{\boldsymbol{\alpha}}$ on $S \subseteq [0,1]^n$, we have*

$$\mu^{cav}(S) \leq \mu(S) \leq \mathcal{C}_d^1, \quad \mu^{vex}(S) \leq \mathcal{C}_d^2,$$

If $\mathbf{0}, \mathbb{1} \in S$, then $\mu(S) = \mu^{cav}(S) = \mathcal{C}_d^1$.

The monotonicity of $\mathcal{C}_d^1$ and $\mathcal{C}_d^2$ with respect to d means that convexifying higher degree monomials is likely to produce a greater error. As $d \to \infty$, we have $\mathcal{C}_d^1 \to 1$ and $\mathcal{C}_d^2 \to 1/e$.

The bounds $\mathcal{C}_d^1$ and $\mathcal{C}_d^2$ depend only on the degree of the monomial. The arguments used in proving Theorem 4.0.1 also imply that a family of convex relaxations of $\boldsymbol{x}^{\boldsymbol{\alpha}}$ has error equal to $\mathcal{C}_d^1$. We also guarantee that the convex envelope error bound $\mathcal{C}_d^2$ is tight for the multilinear monomial $m(\boldsymbol{x})$ on $S = [0,1]^n$.

Theorem 4.0.1 has two immediate implications. First, we obtain the error in convexifying a monomial on $[0,1]^n$.

Corollary 4.0.2. $\mu([0,1]^n) = \mathcal{C}_d^1$.

Second, we obtain a additive error bound on polynomial optimization over subsets of $[0,1]^n$. For a polynomial $p(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha}} c_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\alpha}}$, denote

$$L'(p) = \max\left\{ \max_{\boldsymbol{\alpha}:\, c_{\boldsymbol{\alpha}} > 0} c_{\boldsymbol{\alpha}} \mathcal{C}_d^2,\ \max_{\boldsymbol{\alpha}:\, c_{\boldsymbol{\alpha}} < 0} -c_{\boldsymbol{\alpha}} \mathcal{C}_d^1 \right\}.$$

Let $z(\boldsymbol{x}) \equiv \min\{\sum_{\boldsymbol{\alpha}} c_{\boldsymbol{\alpha}} w_{\boldsymbol{\alpha}} \mid (\boldsymbol{x}, w_{\boldsymbol{\alpha}}) \text{ in the convexification of } \boldsymbol{x}^{\boldsymbol{\alpha}}, \forall \boldsymbol{\alpha}\}$ be the lower bound from monomial convexification on $p(\boldsymbol{x})$.

Corollary 4.0.3. *For any $p(\boldsymbol{x})$ with total degree at most m and compact convex $S \subseteq [0,1]^n$,*

$$p(\boldsymbol{x}) - z(\boldsymbol{x}) \leq L'(p) \binom{n+m}{n}.$$

Proof. We have $z(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha}:\ c_{\boldsymbol{\alpha}}>0} c_{\boldsymbol{\alpha}} \operatorname{vex}_{\boldsymbol{\alpha}}(\boldsymbol{x}) + \sum_{\boldsymbol{\alpha}:\ c_{\boldsymbol{\alpha}}<0} c_{\boldsymbol{\alpha}} \operatorname{cav}_{\boldsymbol{\alpha}}(\boldsymbol{x})$, where $\operatorname{vex}_{\boldsymbol{\alpha}}(\bullet)$ represents the convex envelope of $(\bullet)^{\boldsymbol{\alpha}}$ on S, and $\operatorname{cav}_{\boldsymbol{\alpha}}(\bullet)$ represents the concave envelope of $(\bullet)^{\boldsymbol{\alpha}}$ on S. Therefore,

$$p(\boldsymbol{x}) - z(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha}:\ c_{\boldsymbol{\alpha}}>0} c_{\boldsymbol{\alpha}}(\boldsymbol{x}^{\boldsymbol{\alpha}} - \operatorname{vex}_{\boldsymbol{\alpha}}(\boldsymbol{x})) + \sum_{\boldsymbol{\alpha}:\ c_{\boldsymbol{\alpha}}<0} (-c_{\boldsymbol{\alpha}})(\operatorname{cav}_{\boldsymbol{\alpha}}(\boldsymbol{x}) - \boldsymbol{x}^{\boldsymbol{\alpha}}).$$

Applying Theorem 4.0.1 and the construction of $L'(p)$ gives us $p(\boldsymbol{x}) - z(\boldsymbol{x}) \leq L'(p) \sum_{\boldsymbol{\alpha}} 1$. Since $p(\boldsymbol{x})$ has total degree at most m, there are at most $\binom{n+m}{n}$ monomials in $p(\boldsymbol{x})$, leading to the claimed error bound. □

Computing $L'(p)$ may get tedious if $p(\boldsymbol{x})$ has a large number of monomials. A cheaper bound is possible by considering only the largest coefficient in $p(\boldsymbol{x})$.

Corollary 4.0.4. *For any $p(\boldsymbol{x})$ with total degree at most m and compact convex $S \subseteq [0,1^n$,*

$$p(\boldsymbol{x}) - z(\boldsymbol{x}) \leq \max_{\boldsymbol{\alpha}} |c_{\boldsymbol{\alpha}}| \left(1 - \frac{1}{m}\right) m^{\frac{1}{1-m}} \binom{n+m}{n}.$$

Proof. Follows from Corollary 4.0.3 after using $d \leq m$ and $\mathcal{C}_d^1$ being monotone in d. □

The bounds from Theorem 4.0.1, although applicable to arbitrary $S \subseteq [0,1]^n$, can be weak if $\mathbf{0} \in S$ and $\mathbb{1} \notin S$. To emphasize this, we consider a monomial over the standard n-simplex $\Delta_n \equiv \operatorname{conv}\{\mathbf{0}, \mathbf{e}_1, \ldots, \mathbf{e}_n\} = \left\{\boldsymbol{x} \geq \mathbf{0} \mid \sum_{j=1}^n x_j \leq 1\right\}$, and obtain error bounds that depend on not just the degree of the monomial but also the exponent of each variable. These bounds are stronger than the bounds $\mathcal{C}_d^1$ and $\mathcal{C}_d^2$.

Theorem 4.0.5.

$$\mu^{cav}(\Delta_n) \leq \mu(\Delta_n) \leq \frac{(\boldsymbol{\alpha}^{\boldsymbol{\alpha}})^{1/d}}{d} - \frac{\boldsymbol{\alpha}^{\boldsymbol{\alpha}}}{d^d}, \quad \mu^{vex}(\Delta_n) = \frac{\boldsymbol{\alpha}^{\boldsymbol{\alpha}}}{d^d}.$$

The bound for μ^{cav} is tight if and only if the monomial is symmetric, i.e., $\boldsymbol{\alpha} = \alpha_0 \mathbb{1}$ for some $\alpha_0 \in \mathbb{Z}_{\geq 1}$.

4.0.3 Multilinear monomial

Consider the multilinear monomial $m(\boldsymbol{x}) = \prod_{j=1}^n x_j$.

Theorem 4.0.6. *Denote*

$$\mathcal{D}_{r,n} \equiv \max_{i=1,\dots,n-1} \left\{ \left(1 + \frac{i}{n}(r-1)\right)^n - r^i \right\}, \quad \mathcal{E}_{r,n} \equiv 1 + \frac{r^n - 1}{(r-1)} \left[\frac{n-1}{n} \left(\frac{r^n - 1}{n(r-1)} \right)^{\frac{1}{n-1}} - 1 \right].$$

For $m(\boldsymbol{x})$ on $[1,r]^n$,

$$\mu^{cav}([1,r]^n) = \mathcal{E}_{r,n}, \quad \mu^{vex}([1,r]^n) = \mathcal{D}_{r,n}, \quad \mu([1,r]^n) = \max\{\mathcal{D}_{r,n}, \mathcal{E}_{r,n}\}.$$

All bounds are attained only on the line segment between $\mathbb{1}, r\mathbb{1}$.

We conjecture that $\mathcal{D}_{r,n} \leq \mathcal{E}_{r,n}$ for all r, n and provide a strong empirical evidence in support of this claim. We prove this conjecture to be asymptotically true by showing that $\lim_{n\to\infty} \mathcal{D}_{r,n}/\mathcal{E}_{r,n} < 1/e$.

For $S = [-1,1]^n$, we characterize the convex hull in Section 2.1, and show that it has the following errors.

Theorem 4.0.7. *For $m(\boldsymbol{x})$ on $[-1,1]^n$,*

$$\mu^{cav}([-1,1]^n) = \mu^{vex}([-1,1]^n) = \mu([-1,1]^n) = 1 + \left(\frac{n-2}{n}\right)^n.$$

This maximum error is attained at all the 2^n reflections of the point $(\frac{n-2}{n}\mathbb{1}, -1)$.

Taking $n \to \infty$, this error approaches $1 + 1/e^2$ from below.

4.1 Convex Hull and Monomial Relaxation

Before getting into the analysis, we provide, within this section, the relating convex hull forms and necessary polyhedral relaxation in the $(\boldsymbol{x}, y)$-space, for different domain boxes X' (asymmetric) or X (symmetric). A key note is that, for a multilinear polynomial, the convex hull form is always polyhedral; yet for a general monomial as in Section 4.0.2, which is not multilinear, the convex hull form is always *not* polyhedral, so a polyhedral relaxation is used when a linearization technique is implemented. Meanwhile, due to the special structure of a monomial term, we notice that there is a simple scaling argument for both the convex hull form and the worst-case error analysis.

To be more specific, for any vector $\boldsymbol{c} \in \mathbb{R}^n$ of nonzero scalars, suppose we define a set X'_C in terms of X' by scaling the variables $\boldsymbol{x}$ to obtain

$$\begin{aligned} X'_C \equiv \{\boldsymbol{x} \in \mathbb{R}^n : c_j L_j \leq x_j \leq c_j U_j \ \forall\, j \text{ such that } c_j > 0 \\ c_j U_j \leq x_j \leq c_j L_j \ \forall\, j \text{ such that } c_j < 0\}. \end{aligned} \tag{4.4}$$

(We do not consider $c_j = 0$ for any j since then $\boldsymbol{x}^{\boldsymbol{\alpha}}$ would equal 0 for all $\boldsymbol{x} \in X'_C$.) We have the following scaling lemma:

Lemma 4.1.1. *Given any $\boldsymbol{c} \in \mathbb{R}^n$ with $c_j \neq 0$ for all j, and any monomial of the form $\boldsymbol{x}^{\boldsymbol{\alpha}} = \prod_{j=1}^n x_j^{\alpha_j}$, $\mu(X'_C) = |\boldsymbol{c}^{\boldsymbol{\alpha}}|\, \mu(X')$.*

Proof. We firstly notice that there is a nice linear bijection between the graphs of the monomial over the original and the scaled boxes:

$$\begin{aligned} \{(\boldsymbol{x}, y) \in X'_C \times \mathbb{R} : y = \boldsymbol{x}^{\boldsymbol{\alpha}}\} &= \left\{ (\boldsymbol{x}, y) \in X'_C \times \mathbb{R} : y = \boldsymbol{c}^{\boldsymbol{\alpha}} \prod_{j=1}^n \left(\frac{x_j}{c_j}\right)^{\alpha_j} \right\} \\ &= \left\{ \left(c_1 \frac{x_1}{c_1}, \ldots, c_n \frac{x_n}{c_n}, \boldsymbol{c}^{\boldsymbol{\alpha}} \frac{y}{\boldsymbol{c}^{\boldsymbol{\alpha}}} \right) \in X'_C \times \mathbb{R} : \frac{y}{\boldsymbol{c}^{\boldsymbol{\alpha}}} = \prod_{j=1}^n \left(\frac{x_j}{c_j}\right)^{\alpha_j} \right\} \\ &= \{(c_1 \hat{x}_1, \ldots, c_n \hat{x}_n, \boldsymbol{c}^{\boldsymbol{\alpha}} \hat{y}) : \hat{y} = \hat{\boldsymbol{x}}^{\boldsymbol{\alpha}}, \hat{\boldsymbol{x}} \in X'\}. \end{aligned}$$

This is established upon a change of variable $\hat{\boldsymbol{x}} = \left(\frac{x_1}{c_1}, \ldots, \frac{x_n}{c_n}\right)^T, \hat{y} = \frac{y}{\boldsymbol{c}^{\boldsymbol{\alpha}}}$. Since this bijection is an affine map, it is shared between the convex hulls of the monomial over the original and the scaled boxes as well:

$$\operatorname{conv}\{(\boldsymbol{x}, y) \in X'_C \times \mathbb{R} : y = \boldsymbol{x}^{\boldsymbol{\alpha}}\} = \{(c_1 \hat{x}_1, \ldots, c_n \hat{x}_n, \boldsymbol{c}^{\boldsymbol{\alpha}} \hat{y}) : (\hat{\boldsymbol{x}}, \hat{y}) \in \operatorname{conv}\{(\boldsymbol{x}, y) \in X' \times \mathbb{R} : y = \boldsymbol{x}^{\boldsymbol{\alpha}}\}\}.$$

Therefore

$$\begin{aligned} \mu(X'_C) &= \max\ |y - \boldsymbol{x}^{\boldsymbol{\alpha}}| \ \text{ s.t. } (\boldsymbol{x}, y) \in \operatorname{conv}\{(\boldsymbol{x}, y) \in X'_C \times \mathbb{R} : y = \boldsymbol{x}^{\boldsymbol{\alpha}}\} \\ &= \max\ |\boldsymbol{c}^{\boldsymbol{\alpha}} \hat{y} - \boldsymbol{c}^{\boldsymbol{\alpha}} \hat{\boldsymbol{x}}^{\boldsymbol{\alpha}}| \ \text{ s.t. } (\hat{\boldsymbol{x}}, \hat{y}) \in \operatorname{conv}\{(\boldsymbol{x}, y) \in X' \times \mathbb{R} : y = \boldsymbol{x}^{\boldsymbol{\alpha}}\} \\ &= |\boldsymbol{c}^{\boldsymbol{\alpha}}|\, \mu(X'). \end{aligned}$$

□

Now we list the convex hull forms and polyhedral relaxations:

- On $[L_1, U_1] \times [L_2, U_2]$, the convex hull for $y = x_1 x_2$ is given by the McCormick inequalities:

$$\left\{ (x_1, x_2, y) \in \mathbb{R}^3 : \begin{array}{r} U_2 x_1 + U_1 x_2 - U_1 U_2 \leq y \leq L_2 x_1 + U_1 x_2 - U_1 L_2 \\ L_2 x_1 + L_1 x_2 - L_1 L_2 \leq y \leq U_2 x_1 + L_1 x_2 - L_1 U_2 \end{array} \right\}. \tag{4.5}$$

- On $X = [0,1]^n$, the convex hull for $y = \prod_{j=1}^n x_j$ is given by the Glover and Woolsey inequalities:

$$\left\{ \begin{array}{rl} (\boldsymbol{x}, y) \in \mathbb{R}^{n+1} : & \\ & y \leq x_j \; \forall \, j = 1, \ldots, n \\ & x_j \leq 1 \; \forall \, j = 1, \ldots, n \\ \sum_{j=1}^n x_j - (n-1) \leq & y \\ 0 \leq & y \end{array} \right\}. \tag{4.6}$$

- On $X = [0,1]^n$, the convex hull for $y = \boldsymbol{x}^{\boldsymbol{\alpha}}$ is not polyhedral, yet upon duplicating variables with the Glover and Woolsey inequalities of (4.6), we obtain the following polyhedral relaxation:

$$\left\{ \begin{array}{rl} (\boldsymbol{x}, y) \in \mathbb{R}^{n+1} : & \\ & y \leq x_j \; \forall \, j = 1, \ldots, n \\ & x_j \leq 1 \; \forall \, j = 1, \ldots, n \\ \sum_{j=1}^n \alpha_j x_j - (d-1) \leq & y \\ 0 \leq & y \end{array} \right\}. \tag{4.7}$$

  where $d = \sum_{j=1}^n \alpha_j$ is the degree of the monomial. Based on our analysis from Chapter 2, its "plot" is identical to the "plot" of the multilinear monomial, and hence the over-estimators here give the concave envelope, and the inequality $y \geq 0$ gives a facet. Therefore (4.7) is not strong in the sense that, to obtain the convex hull, there needs to be countably many inequalities other than $y \geq \sum_{j=1}^n \alpha_j x_j - (d-1)$, which itself is tight, because it is the tangent plane through the point $\mathbb{1}$ in the $(\boldsymbol{x}, y)$-space. To conclude this item, any other tight underestimating inequality should not go through $\mathbb{1}$.

- We see from the proof of Lemma 4.1.1, that upon scaling the domain box of a monomial, the convex hull form will be scaled accordingly. Given $y = \prod_{j=1}^{n} x_j$ on the set X' having *either* $L_j = 0$ *or* $U_j = 0$ for each j, and denoting $I^1 = \{j : L_j = 0\}$ and $I^2 = \{j : U_j = 0\}$, we have its convex hull form upon utilizing the Glover and Woolsey inequalities of (4.6):

$$\left\{ \begin{array}{rl} (\boldsymbol{x}, y) \in \mathbb{R}^{n+1} : & \\ (-1)^{|I^2|} y \leq & \left(\prod_{i \neq j} \Delta_i \right) x_j \; \forall \; j \in I^1 \\ (-1)^{|I^2|} y \leq & \left(\prod_{i \neq j} \Delta_i \right) (-x_j) \; \forall \; j \in I^2 \\ x_j \leq & U_j \; \forall \; j \in I^1 \\ L_j \leq & x_j \; \forall \; j \in I^2 \\ \sum_{j \in I^1} \left(\prod_{i \neq j} \Delta_i \right) x_j - & \sum_{j \in I^2} \left(\prod_{i \neq j} \Delta_i \right) x_j - (n-1) \left(\prod_{i} \Delta_i \right) \leq (-1)^{|I^2|} y \\ 0 \leq & (-1)^{|I^2|} y \end{array} \right\} \tag{4.8}$$

  where $\Delta_j = (U_j - L_j)$ for all j.

- On $X = [1, r]^n$, $y = \prod_{j=1}^{n} x_j$ is an SMP that is strictly supermodular, i.e., it has a strictly convex "plot". Utilizing the known results from the literature or our Theorem 2.3.1, its convex hull is given by the following inequalities:

$$y \leq \min_{n\text{-permutation } \sigma} \sum_{j=1}^{n} r^{j-1} x_{\sigma(j)} - \sum_{j=1}^{n-1} r^j, \tag{4.9}$$

$$y \geq \max_{1 \leq i \leq n} r^{i-1} \left(\sum_{j=1}^{n} x_j - (n-i) - r(i-1) \right), \tag{4.10}$$

  and of course $\boldsymbol{x} \in X$.

- On $X = [-1, 1]^n$, $y = \prod_{j=1}^{n} x_j$ is an SMP that has an alternating "plot." Rearranging the convex hull form (2.15) derived in Section 2.3.2 to express it explicitly in terms of y, we have $-1 \leq y \leq 1$, together with the following:

$$\max_{S \subseteq N : |S| \text{ even}} \sum_{j \in N \setminus S} x_j - \sum_{j \in S} x_j - (n-1) \leq y \leq \min_{T \subseteq N : |T| \text{ even}} \sum_{j \in N \setminus T} x_j - \sum_{j \in T} x_j + (n-1) \tag{4.11}$$

if n is odd, or

$$\max_{S\subseteq N:|S| \text{ even}} \sum_{j\in N\setminus S} x_j - \sum_{j\in S} x_j - (n-1) \le y \le \min_{T\subseteq N:|T| \text{ odd}} \sum_{j\in N\setminus T} x_j - \sum_{j\in T} x_j + (n-1) \quad (4.12)$$

if n is even.

4.2 Proof of Theorem 4.0.1, where $X' = X = [0,1]^n$

The proofs of the theorems listed at the beginning of the Chapter share a common feature: the monomial as a hypersurface in the $(\boldsymbol{x}, y)$-space is squeezed between the over-estimator and under-estimator of the convex hull form; thus we analyze μ^{cav} of (4.2) and μ^{vex} of (4.3) separately, and compare these worst-case errors to conclude μ of (4.1).

Proof of Theorem 4.0.1. Notice that by definition $\mu^{\text{cav}}(\bullet)$, $\mu^{\text{vex}}(\bullet)$ and $\mu(\bullet)$ are "monotone," i.e., $A \subseteq B \subseteq \mathbb{R}^n$ implies $\mu^{\text{cav}}(A) \le \mu^{\text{cav}}(B)$, etc. It suffices to show that $\mu^{\text{cav}}\left([0,1]^n\right) = \mathcal{C}_d^1$ and that $\mu^{\text{vex}}\left([0,1]^n\right) \le \mathcal{C}_d^2$. Here we make use of the relaxation (4.7).

For $\mu^{\text{cav}}\left([0,1]^n\right)$ regarding the concave over-estimator, for any $(\boldsymbol{x}, y)$ that satisfies (4.7), we can provide the following string of upper bounding inequalities:

$$\begin{aligned} y - \boldsymbol{x}^{\boldsymbol{\alpha}} &\le \min_{1\le j\le n} x_j - \boldsymbol{x}^{\boldsymbol{\alpha}} && (4.13)\\ &\le (\boldsymbol{x}^{\boldsymbol{\alpha}})^{\frac{1}{d}} - \boldsymbol{x}^{\boldsymbol{\alpha}} && (4.14)\\ &\le \frac{d-1}{d}\left(\frac{1}{d}\right)^{\frac{1}{d-1}} && (4.15)\\ &= \mathcal{C}_d^1. \end{aligned}$$

(4.13) is because of the over-estimator of (4.7); (4.14) is trivial since all entries of $\boldsymbol{x}$ are in $[0,1]$. (4.15) is not hard to obtain if we treat $\boldsymbol{x}^{\boldsymbol{\alpha}}$ as another variable $t \in [0,1]$ and compute the maximum value: $\max_{t\in[0,1]} t^{\frac{1}{d}} - t$ is attained as $\mathcal{C}_d^1$ at its only critical point $t = \left(\frac{1}{d}\right)^{\frac{d}{d-1}} \in (0,1)$. $\mu^{\text{cav}}\left([0,1]^n\right) = \mathcal{C}_d^1$ forces the above inequalities to hold as equalities, thus the worst-case error is attained at $(\boldsymbol{x}, y) = (\left(\frac{1}{d}\right)^{\frac{1}{d-1}} \mathbb{1}, \left(\frac{1}{d}\right)^{\frac{1}{d-1}})$.

For $\mu^{\text{vex}}\left([0,1]^n\right)$ regarding the convex under-estimator, for any $(\boldsymbol{x}, y)$ that satisfies (4.7),

we can provide the following string of upper bounding inequalities:

$$\begin{aligned}
\boldsymbol{x}^{\boldsymbol{\alpha}} - y &\leq \boldsymbol{x}^{\boldsymbol{\alpha}} - \max\{0, \sum_{j=1}^{n} \alpha_j x_j - (d-1)\} && (4.16)\\
&= \min\{\boldsymbol{x}^{\boldsymbol{\alpha}}, \boldsymbol{x}^{\boldsymbol{\alpha}} - \sum_{j=1}^{n} \alpha_j x_j + (d-1)\}\\
&\leq \min\{\boldsymbol{x}^{\boldsymbol{\alpha}}, \boldsymbol{x}^{\boldsymbol{\alpha}} - d\,(\boldsymbol{x}^{\boldsymbol{\alpha}})^{\frac{1}{d}} + (d-1)\} && (4.17)\\
&\leq \left(\frac{d-1}{d}\right)^{d} && (4.18)\\
&= \mathcal{C}_d^2.
\end{aligned}$$

(4.16) is because of the definition of (4.7); (4.17) is true by the inequality of arithmetic and geometric means. (4.18) is not hard to obtain if we treat $\boldsymbol{x}^{\boldsymbol{\alpha}}$ as another variable $t \in [0,1]$ and compute the maximum value: $\min\{t, t - dt^{\frac{1}{d}} + (d-1)\}$ is increasing on $[0, \left(\frac{d-1}{d}\right)^d]$ and then decreasing on $[\left(\frac{d-1}{d}\right)^d, 1]$, and hence reaches its maximum $\mathcal{C}_d^2$ at $t = \mathcal{C}_d^2$. For this relaxation (4.7) the worst-case error is attained if and only if the above inequalities hold as equalities, i.e., at $(\boldsymbol{x}, y) = (\frac{d-1}{d}\mathbb{1}, 0)$.

Finally, we show that $\mathcal{C}_d^1 \geq \mathcal{C}_d^2$ for $d \geq 2$, with equality only at $d = 2$:

$$\frac{d-1}{d}\left(\frac{1}{d}\right)^{\frac{1}{d-1}} \geq \left(\frac{d-1}{d}\right)^{d} \Leftrightarrow \left(\frac{d}{d-1}\right)^{(d-1)^2} \geq d,$$

while the latter is true because:

$$\left(1 + \frac{1}{d-1}\right)^{(d-1)^2} \geq 1 + (d-1)^2 \frac{1}{d-1} = d \qquad (4.19)$$

by the binomial expansion, and when $d \geq 3$, the inequality is strict. □

Remark 6. (i) *The inequality of arithmetic and geometric means plays a very important role in the analysis of convex hull errors, as it guarantees that the proposed worst-case error can be achieved.*

(ii) *Since the convex under-estimator of* (4.7) *is not tight,* $\mathcal{C}_d^2$ *is not a tight bound for any stronger relaxation that cuts off* $\frac{d-1}{d}\mathbb{1}$. *To obtain stronger bounds for* $\mu^{vex}([0,1]^n)$, *one may need a stronger relaxation form or even the strongest form — the convex envelope for the monomial. The latter approach is highly challenging because the form is unknown in general, and the*

associating optimization problem, as in the previous proof, is hard to solve analytically even for special known cases. The former approach is also challenging because the cuts may also make the associating optimization problem hard to solve analytically, and *the inequality of arithmetic and geometric means may not be applicable to provide a point that achieves the proposed error. There is no clear utility of better bounds other than* $\mathcal{C}_d^2$ *— it is already much smaller than* $\mathcal{C}_d^1$.

We conclude the section by proving Theorem 4.0.5. The situation is opposite for the analysis over the standard n-simplex Δ_n, since the concave over-estimator is not tight, while the convex under-estimator is.

Proof of Theorem 4.0.5. We still use the relaxation (4.7) for the analysis. Notice that on Δ_n the over-estimator of (4.7) is not tight, because $\mathbb{1}$ is not in the domain. Yet based on Chapter 2, the convex envelope for the monomial on Δ_n is exactly $y = 0$.

For $\mu^{\text{cav}}(\Delta_n)$ regarding the concave over-estimator, following exactly the same steps (4.13), (4.14) and letting $t = \boldsymbol{x}^{\boldsymbol{\alpha}}$, we need to update the maximum value of t :

$$\boldsymbol{x}^{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^{\boldsymbol{\alpha}} \prod_{j=1}^{n} \left(\frac{x_j}{\alpha_j}\right)^{\alpha_j} \leq \boldsymbol{\alpha}^{\boldsymbol{\alpha}} \left(\frac{\sum_{j=1}^{n} x_j}{d}\right)^{d} \leq \boldsymbol{\alpha}^{\boldsymbol{\alpha}} \left(\frac{1}{d}\right)^{d}, \tag{4.20}$$

again, by the inequality of arithmetic and geometric means, and the definition of the simplex. We show next that $\boldsymbol{\alpha}^{\boldsymbol{\alpha}}/d^d \leq \left(\frac{1}{d}\right)^{\frac{d}{d-1}}$, the critical point from the previous proof, hence by the monotonicity of $t^{\frac{1}{d}} - t$, $\mu^{\text{cav}}(\Delta_n) \leq \frac{(\boldsymbol{\alpha}^{\boldsymbol{\alpha}})^{1/d}}{d} - \frac{\boldsymbol{\alpha}^{\boldsymbol{\alpha}}}{d^d}$: For fixed integers $2 \leq n \leq d$, it is easy to argue that

$$\max_{\boldsymbol{\alpha}} \left\{ \boldsymbol{\alpha}^{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \mathbb{Z}^n_{\geq 1}, \sum_{j=1}^{n} \alpha_j = d \right\} \;=\; (d-n+1)^{d-n+1},$$

using the convexity of $\xi \in (0, \infty) \mapsto \xi \ln \xi$. Therefore for fixed d, the "maximum possible value" of $\boldsymbol{\alpha}^{\boldsymbol{\alpha}}$ is achieved with $n = 2$ and is equal to $(d-1)^{d-1}$. Thus, $\boldsymbol{\alpha}^{\boldsymbol{\alpha}}/d^d \leq (d-1)^{d-1}/d^d$. Now $(d-1)^{d-1}/d^d \leq \left(\frac{1}{d}\right)^{\frac{d}{d-1}}$ is equivalent to $(d-1)^{(d-1)^2} \leq d^{d(d-2)}$, which is already shown by (4.19).

According to the equality conditions of (4.13) and (4.14), the proposed bound for μ^{cav} is attained only at $(\boldsymbol{x}, y) = (\frac{(\boldsymbol{\alpha}^{\boldsymbol{\alpha}})^{1/d}}{d}\mathbb{1}, \frac{(\boldsymbol{\alpha}^{\boldsymbol{\alpha}})^{1/d}}{d})$, under the condition that $\frac{(\boldsymbol{\alpha}^{\boldsymbol{\alpha}})^{1/d}}{d}\mathbb{1} \in \Delta_n$. However,

$$\boldsymbol{\alpha}^{\boldsymbol{\alpha}} \geq (d/n)^d,$$

which is obtained by applying Jensen's inequality to the convex function $\xi \in (0,\infty) \mapsto \xi \log \xi$ with the n points being $\xi_i = \alpha_i \ \forall i$ and the convex combination weights being equal to $1/n$. The equality condition is due to $\xi \log \xi$ being strictly convex.

For $\mu^{\text{vex}}(\Delta_n)$ regarding the convex under-estimator, as the only facet is $y = 0$, it suffices to compute $\max_{\boldsymbol{x} \in \Delta_n} \boldsymbol{x}^{\boldsymbol{\alpha}}$, which is again $\boldsymbol{\alpha}^{\boldsymbol{\alpha}}/d^d$ given by (4.20), and is attained at $(\boldsymbol{x}, y) = (\alpha_1/d, \ldots, \alpha_n/d, 0)$.

Finally, as a comparison between the over-estimator and under-estimator bounds, we show that $\frac{(\boldsymbol{\alpha}^{\boldsymbol{\alpha}})^{1/d}}{d} - \frac{\boldsymbol{\alpha}^{\boldsymbol{\alpha}}}{d^d} \geq \frac{\boldsymbol{\alpha}^{\boldsymbol{\alpha}}}{d^d}$:

$$2^d \left(\boldsymbol{\alpha}^{\boldsymbol{\alpha}}\right)^{d-1} \leq 2^d (d-1)^{(d-1)^2} \leq 2^d d^{d(d-2)} \leq d^{d(d-1)},$$

based on the previous calculation for the "maximum possible value" of $\boldsymbol{\alpha}^{\boldsymbol{\alpha}}$, implementation of (4.19), and the fact that $d \geq 2$. □

4.3 Proof of Theorem 4.0.6, where $X' = X = [1, r]^n$

Although the proof shares a common flavor with the previous proofs, and the worst-errors appear on the segment between $\mathbb{1}$ and $r\mathbb{1}$ as well, the argument and analytical bounds are much more complex and the error comparison is harder due to the complexity of the envelope forms (4.9) and (4.10).

Proof of Theorem 4.0.6. For $\mu^{\text{cav}}([1, r]^n)$ regarding the concave over-estimator, given any $(\boldsymbol{x}, y)$ that satisfies (4.9),

$$y - \prod_{j=1}^{n} x_j \leq \min_{n\text{-permutation } \sigma} \sum_{j=1}^{n} r^{j-1} x_{\sigma(j)} - \sum_{j=1}^{n-1} r^j - \prod_{j=1}^{n} x_j,$$

and we wish to maximize the right-hand side subject to all $\boldsymbol{x}$. Notice that the terms subtracted from the right-hand side are invariant under permutations in $\boldsymbol{x}$; due to the same logic as in the proof of Theorem 2.2.1, $\sum_{j=1}^{n} r^{j-1} x_{\sigma(j)}$ subject to all n-permutations is minimized for the n-permutation σ that has $x_{\sigma(1)} \geq \ldots \geq x_{\sigma(n)}$. This means all permutations of a fixed $\boldsymbol{x}$ leads to the same right-hand side value, i.e., the right-hand side as a function in $\boldsymbol{x}$ is symmetric. Thus, without loss of generality,

we can assume $x_1 \leq \ldots \leq x_n$, and the key term in the right-hand side becomes

$$\min_{n\text{-permutation } \sigma} \sum_{j=1}^{n} r^{j-1} x_{\sigma(j)} = \sum_{j=1}^{n} r^{n-j} x_j.$$

Then we concentrate on the maximizer of

$$\max_{1 \leq x_1 \leq \ldots \leq x_n \leq r} \left\{ \sum_{j=1}^{n} r^{n-j} x_j - \prod_{j=1}^{n} x_j \right\} - \sum_{j=1}^{n-1} r^j,$$

and claim that a maximizer $\boldsymbol{x}$ cannot have $x_i < x_{i+1}$ for any $i \in \{1, \ldots n-1\}$. Otherwise, x_i can be adjusted larger yet still stay feasible, but, as a maximizer, the partial derivative with respect to x_i should not be positive; similarly x_{i+1} can be adjusted less yet still stay feasible, but, as a maximizer, the partial derivative with respect to x_{i+1} should not be negative; i.e., one must have

$$r^{n-i} - \prod_{j \neq i} x_j \leq 0, \quad r^{n-i-1} - \prod_{j \neq i+1} x_j \geq 0,$$

which leads to the following chain of inequalities:

$$r^{n-i} \leq r^{n-i} x_i \leq \prod_{j=1}^{n} x_j \leq r^{n-i-1} x_{i+1} \leq r^{n-i-1} r.$$

This means each inequality above must hold with equality, hence $x_1 = \ldots = x_i = 1, x_{i+1} = \ldots = x_n = r$, but at this $\boldsymbol{x}$ the objective is 0, which is clearly not a maximizer. Now that a maximizer has all entries equal, to some $t \in [1, r]$, we have the desired bound

$$\max_{1 \leq t \leq r} \left\{ t \sum_{j=1}^{n} r^{n-j} - t^n \right\} - \sum_{j=1}^{n-1} r^j = \mathcal{E}_{r,n},$$

which is attained at $(\boldsymbol{x}, y) = \left(\left(\frac{r^n - 1}{n(r-1)} \right)^{\frac{1}{n-1}} \mathbb{1}, \frac{r^n - 1}{r-1} \left[\left(\frac{r^n - 1}{n(r-1)} \right)^{\frac{1}{n-1}} - 1 \right] + 1 \right)$.

For $\mu^{\text{vex}}([1, r]^n)$ regarding the convex under-estimator, given any $(\boldsymbol{x}, y)$ that satisfies (4.10),

we can provide the following string of upper bounding inequalities:

$$\begin{aligned}
\prod_{j=1}^{n} x_j - y &\leq \prod_{j=1}^{n} x_j - \max_{1\leq i\leq n} r^{i-1}\left(\sum_{j=1}^{n} x_j - (n-i) - r(i-1)\right) && (4.21)\\
&= \prod_{j=1}^{n} x_j + \min_{1\leq i\leq n} r^{i-1}\left(-\sum_{j=1}^{n} x_j + (n-i) + r(i-1)\right)\\
&\leq \prod_{j=1}^{n} x_j + \min_{1\leq i\leq n} r^{i-1}\left(-n\left(\prod_{j=1}^{n} x_j\right)^{\frac{1}{n}} + (n-i) + r(i-1)\right) && (4.22)\\
&\leq \max_{1\leq i\leq n}\left(1+\frac{i-1}{n}(r-1)\right)^{n} - r^{i-1} && (4.23)\\
&= \mathcal{D}_{r,n}.
\end{aligned}$$

(4.21) is true because of the definition of (4.10); (4.22) is true by the inequality of arithmetic and geometric means. (4.23) is not hard to obtain if we treat $\left(\prod_{j=1}^{n} x_j\right)^{\frac{1}{n}}$ as another variable $t \in [1, r]$ and observe that

1. the function

$$t \in [1, r] \mapsto t^n + \min_{1\leq i\leq n} r^{i-1}\left(-nt + (n-i) + r(i-1)\right)$$

   is a convex function plus a piecewise linear function;

2. the i-th linear piece $r^{i-1}\left(-nt + (n-i) + r(i-1)\right)$ is the "lowest" on $t \in [1+\frac{i-1}{n}r, 1+\frac{i}{n}r]$, the same reason as in (4.10), that the i-th facet is the "highest" over the polytope $\{\boldsymbol{x} \in [1, r]^n : (i-1)r + (n-i+1) \leq \sum_{j=1}^{n} x_j \leq ir + (n-i)\}$;

3. therefore this function on $[1 + \frac{i-1}{n}r, 1 + \frac{i}{n}r]$ is exactly $t^n + r^{i-1}\left(-nt + (n-i) + r(i-1)\right)$, which is convex and hence attains maximum on the endpoints.

Due to (4.22), by the inequality of arithmetic and geometric means, this worst-case error must attain on the segment between $\mathbb{1}, r\mathbb{1}$. □

Notice that the range of the monomial $\prod_{j=1}^{n} x_j$ on the $[1, r]^n$ box is of a length $r^n - 1$, so asymptotically, as $n \to \infty$, $\frac{\mathcal{E}_{r,n}}{r^n-1} \to 1$, based on the fact that $\left(\frac{r^n-1}{n(r-1)}\right)^{\frac{1}{n-1}} \to r$. The comparison between $\mathcal{E}_{r,n}$ and $\mathcal{D}_{r,n}$ faces the following challenges:

1. the expression of $\mathcal{E}_{r,n}$ is explicit, yet complex;

2. the expression of $\mathcal{D}_{r,n}$ is implicit, which makes the asymptotic analysis hard to perform;

3. even if both expressions are explicit, it remains a challenge to theoretically perform the comparison as in the previous section. A numerical comparison is then desired.

In order to handle item 2, we analyze $\mathcal{D}_{r,n}$ deeper:

$$\max_{1\le i\le n}\left(1+\frac{i-1}{n}(r-1)\right)^n - r^{i-1} \le \max_{0\le t\le 1}(1+t(r-1))^n - r^{nt},$$

with latter function obtained from substituting $\frac{i-1}{n}$ with t, it is natural to analyze the latter function on $[0,1]$. We have the following result as a more explicit expression for $\mathcal{D}_{r,n}$:

Proposition 4.3.1. *Denote W_{-1} as the lower branch of the Lambert W function.*

$$\mathcal{D}_{r,n} = \max\left\{\left(1+\frac{i_0-1}{n}(r-1)\right)^n - r^{i_0-1}, \left(1+\frac{i_0}{n}(r-1)\right)^n - r^{i_0}\right\},$$

where

$$i_0 = \left\lceil -\frac{n}{r-1} - \frac{n-1}{\ln r} W_{-1}\left(-\frac{n}{n-1}\left(\frac{\ln r}{r-1}\right)^{\frac{n}{n-1}} r^{-\frac{n}{(n-1)(r-1)}}\right)\right\rceil.$$

Proof. We know that $(1+t(r-1))^n - r^{nt}, t\in[0,1]$ has at least one critical point due to Rolle's Theorem. To identify the critical point(s) one can set its derivative equal to 0 with respect to t:

$$(1+t(r-1))^{n-1}(r-1) = r^{tn}\ln r,$$

After a series of transformations one could obtain the following:

$$-\frac{n\ln r}{n-1}\left(t+\frac{1}{r-1}\right)e^{-\frac{n\ln r}{n-1}\left(t+\frac{1}{r-1}\right)} = -\frac{n}{n-1}\left(\frac{\ln r}{r-1}\right)^{\frac{n}{n-1}} r^{-\frac{n}{(n-1)(r-1)}}.$$

The existence of a critical number indicates that there must be some t that satisfies the above equality. And based on our knowledge of the function $y = xe^x$, since the right-hand side of the above is in $[-e^{-1}, 0)$, the above equality may have two distinct solutions. To see that there is exactly one solution on $[0,1]$ that is given by the lower branch of the Lambert W function, one just needs to verify that when $t=0$, the left-hand side is strictly less than the right-hand side, i.e.,

$$-\frac{n\ln r}{(n-1)(r-1)} r^{-\frac{n}{(n-1)(r-1)}} < -\frac{n}{n-1}\left(\frac{\ln r}{r-1}\right)^{\frac{n}{n-1}} r^{-\frac{n}{(n-1)(r-1)}}.$$

Then simply apply the Lambert W function to obtain the unique critical number $t^* = -\frac{1}{r-1} - \frac{n-1}{n\ln r} W_{-1}\left(-\frac{n}{n-1}(\frac{\ln r}{r-1})^{\frac{n}{n-1}} r^{-\frac{n}{(n-1)(r-1)}}\right) \in (0,1)$. Since the critical point on $[0,1]$ is unique, $(1+t(r-1))^n - r^{nt}$ increases on $(0,t^*)$, decreases on $(t^*,1)$, and thus $\mathcal{D}_{r,n}$ is obtained either at $i = \lceil nt^* \rceil$ or $i = \lceil nt^* \rceil + 1$. □

Although it is not difficult to compute and compare $\mathcal{D}_{r,n}$ and $\mathcal{E}_{r,n}$ computationally in any instance one might encounter, it is still theoretically challenging to prove the conjecture that $\mathcal{D}_{r,n} \leq \mathcal{E}_{r,n}$: one must be very careful on enlargement of expression and manipulation of bounds, since when $n = 2$, the concave envelope error and the convex envelope error both equal $\frac{(U_1-L_1)(U_2-L_2)}{4} = \frac{(r-1)^2}{4}$ — there is no gap! However, we have the following theoretical bound to help us compare asymptotically.

Proposition 4.3.2. $\mathcal{D}_{r,n} \leq \max\left\{\left(1+\frac{n-1}{n}(r-1)\right)^n - r^{n-1}, r^n(\frac{\ln r}{r-1})^{\frac{n}{n-1}} - r^{n-1}\right\}$.

Proof. From the analysis within the proof of Proposition 4.3.1, we know that on $[0,1]$ the critical point t^* is unique, and it satisfies

$$(1+t^*(r-1))^{n-1} = r^{t^* n}\frac{\ln r}{r-1}.$$

If $t^* \geq \frac{n-1}{n}$, then clearly $\mathcal{D}_{r,n} = \left(1+\frac{n-1}{n}(r-1)\right)^n - r^{n-1}$; however, if $t^* \leq \frac{n-1}{n}$, one can have that

$$\begin{aligned}
\mathcal{D}_{r,n} &\leq \max_{0\leq t\leq 1} (1+t(r-1))^n - r^{nt} \\
&= (1+t^*(r-1))^n - r^{nt^*} \\
&= r^{\frac{n^2 t^*}{n-1}}\left(\frac{\ln r}{r-1}\right)^{\frac{n}{n-1}} - r^{nt^*} \\
&= r^{nt^*}\left[r^{\frac{nt^*}{n-1}}\left(\frac{\ln r}{r-1}\right)^{\frac{n}{n-1}} - 1\right] \\
&\leq r^{n-1}\left[r\left(\frac{\ln r}{r-1}\right)^{\frac{n}{n-1}} - 1\right].
\end{aligned}$$

For the last inequality, because $\mathcal{D}_{r,n}$ is positive, $r^{\frac{nt^*}{n-1}}(\frac{\ln r}{r-1})^{\frac{n}{n-1}} - 1$ is forced to be positive, and there will be no problem enlarging t^* all the way to $\frac{n-1}{n}$. □

Remark 7. *In some scenarios, $r(\frac{\ln r}{r-1})^{\frac{n}{n-1}} - 1$ is negative, in which case it must be that $t^* \geq \frac{n-1}{n}$. Now we can compute the following and provide an asymptotic bound for $\mathcal{D}_{r,n}/r^n$:*

(i) $\left(1-\frac{1}{n}+\frac{1}{nr}\right)^n - r^{-1} \to e^{\frac{1}{r}-1} - \frac{1}{r} \leq e^{-1} < 0.37$;

(ii) $(\frac{\ln r}{r-1})^{\frac{n}{n-1}} - r^{-1} \leq \frac{\ln r}{r-1} - \frac{1}{r} < 0.22$.

Example 3. *Consider the monomial* $\prod_{j=1}^n x_j$ *in* $n = 3$ *variables when* $L_1 = L_2 = L_3 = 1$ *and* $U_1 = U_2 = U_3 = 4$ *in* X' *of* (1.4) *so that the conditions of Theorem 4.0.6 are satisfied with* $r = 4$. *Theorem 4.0.6 gives* $\mathcal{D}_{r,n} = 11$, *which is only realized at* $(\boldsymbol{x}, y) = (3, 3, 3, 16)$; $\mathcal{E}_{r,n} = 14\sqrt{7} - 20 \approx 17.04$, *which is only realized at the* $(\boldsymbol{x}, w) = (\sqrt{7}, \sqrt{7}, \sqrt{7}, 21\sqrt{7} - 20)$; *and the error* $\mu([1, 4]^3)$ *of* (4.1) *is* $\mathcal{E}_{r,n} = 14\sqrt{7} - 20$.

4.4 Proof of Theorem 4.0.7, where $X' = X = [-1, 1]^n$

Proof of Theorem 4.0.7. We first make use of the reflection symmetry in the sets G of (1.1) and $\text{conv}(G)$ — the "graph" and the convex hull of the "graph" of the monomial over domain box $X = [-1, 1]^n$ — to simplify the problem.

Define the sign flipping map τ_j for each $j = 1, \ldots, n$ and τ as follows: for any $(\boldsymbol{x}, y) \in [-1, 1]^{n+1}$, $\tau_j(\boldsymbol{x}) = \boldsymbol{x} - 2x_j\mathbf{e}_j$, and $\tau(y) = -y$. τ_j simply flip the sign of entry j in a vector, and τ simply flip the sign of a real number. For any j, it is clear that $(\boldsymbol{x}, y) \in G$ if and only if $(\tau_j(\boldsymbol{x}), \tau(y)) \in G$; a simple point-set argument can carry this equivalence into $\text{conv}(G)$, that $(\boldsymbol{x}, y) \in \text{conv}(G)$ if and only if $(\tau_j(\boldsymbol{x}), \tau(y)) \in \text{conv}(G)$.

Notice that the error measure $h(\boldsymbol{x}, y) \equiv \left| y - \prod_{j=1}^n x_j \right|$ is preserved between $(\boldsymbol{x}, y)$ and $(\tau_j(\boldsymbol{x}), \tau(y))$, which means that to study the approximation error at the point $(\boldsymbol{x}, y) \in \text{conv}(G)$, it is equivalent to flip one entry in $\boldsymbol{x}$ and flip y and study the error at the point $(\tau_j(\boldsymbol{x}), \tau(y)) \in \text{conv}(G)$. At most n flips will result in a $(n+1)$-dimensional vector with its first n entries being nonnegative. In other words,

$$\mu([-1, 1]^n) = \max \left\{ \left| y - \prod_{j=1}^n x_j \right| : (\boldsymbol{x}, y) \in \text{conv}(G), \boldsymbol{x} \geq \mathbf{0} \right\},$$

meaning that we only need to consider without loss of generality, $\boldsymbol{x} \geq \mathbf{0}$ when computing the convex hull error.

When n is odd, for $\mu^{\text{cav}}([-1, 1]^n)$ regarding the concave over-estimator, for any $(\boldsymbol{x}, y)$ that

satisfies (4.11) and $y \leq 1$, we can provide the following string of upper bounding inequalities:

$$\begin{aligned} y - \prod_{j=1}^{n} x_j &\leq \min \left\{ \min_{T \subseteq N, |T| \text{ even}} \sum_{j \in N \setminus T} x_j - \sum_{j \in T} x_j + (n-1), 1 \right\} - \prod_{j=1}^{n} x_j \qquad (4.24) \\ &\leq 1 - \prod_{i=1}^{n} x_i \leq 1. \end{aligned}$$

For $\mu^{\text{vex}}([-1,1]^n)$ regarding the convex under-estimator, for any $(\boldsymbol{x}, y)$ that satisfies (4.11) and $y \geq -1$, we can provide the following string of upper bounding inequalities:

$$\begin{aligned} \prod_{j=1}^{n} x_j - y &\leq \prod_{j=1}^{n} x_j - \max \left\{ \max_{S \subseteq N, |S| \text{ even}} \sum_{j \in N \setminus S} x_j - \sum_{j \in S} x_j - (n-1), -1 \right\} \qquad (4.25) \\ &= \prod_{j=1}^{n} x_j - \max \left\{ \sum_{j=1}^{n} x_j - (n-1), -1 \right\} \\ &= \prod_{j=1}^{n} x_j + \min \left\{ -\sum_{j=1}^{n} x_j + (n-1), 1 \right\} \\ &\leq \prod_{j=1}^{n} x_j + \min \left\{ -n \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}} + (n-1), 1 \right\} \qquad (4.26) \\ &\leq 1 + \left(\frac{n-2}{n} \right)^n. \qquad (4.27) \end{aligned}$$

(4.24) and (4.25) are true because of the definition of (4.11) and $-1 \leq y \leq 1$. (4.26) is true by the inequality of arithmetic and geometric means. (4.27) is not hard to obtain if we treat $\prod_{i=1}^{n} x_i$ as another variable $t \in [0,1]$ and compute the maximum value: $t + \min \left\{ -nt^{\frac{1}{n}} + (n-1), 1 \right\}$ is increasing on $\left[0, \left(\frac{n-2}{n}\right)^n\right]$ and then decreasing on $\left[\left(\frac{n-2}{n}\right)^n, 1\right]$, hence it reaches its maximum $1 + \left(\frac{n-2}{n}\right)^n$ at $t = \left(\frac{n-2}{n}\right)^n$.

When n is even, we can perform an almost identical argument as the above two strings of inequalities, except in (4.24) in which $|T|$ has to be odd due to the over-estimator within (4.12). The same conclusion follows. Therefore, regardless of the parity of n, within $\boldsymbol{x} \geq \boldsymbol{0}$, the worst-case error is attained if and only if the second string of inequalities hold as equalities, i.e., at $(\boldsymbol{x}, y) = (\frac{n-2}{n}\mathbb{1}, -1)$. The reflection symmetry of the problem established at the beginning of the proof indicates that this worst-case error is realized at all 2^n reflected copies of $(\boldsymbol{x}, y) = (\frac{n-2}{n}\mathbb{1}, -1)$, which must satisfy $y \prod_{j=1}^{n} x_j < 0$.

□

This error we obtain is greater than 1, and it asymptotically tends to $1 + \frac{1}{e^2}$.

Example 4. *Consider the monomial* $\prod_{j=1}^{n} x_j$ *in* $n = 3$ *variables when* $L_1 = -U_1 = -2$, $L_2 = -U_2 = -3$, *and* $L_4 = -U_4 = -4$ *in* X' *of* (1.4). *Theorem 4.0.7 and Lemma 4.1.1 gives the error* $\mu(X')$ *of* (4.1) *as* $24\frac{8}{9}$, *and this error occurs at only the eight points* $(\boldsymbol{x}, y) = (\frac{2}{3}, 1, \frac{4}{3}, -24)$, $(-\frac{2}{3}, 1, \frac{4}{3}, 24)$, $(\frac{2}{3}, -1, \frac{4}{3}, 24)$, $(\frac{2}{3}, 1, -\frac{4}{3}, 24)$, $(-\frac{2}{3}, -1, \frac{4}{3}, -24)$, $(-\frac{2}{3}, 1, -\frac{4}{3}, -24)$, $(\frac{2}{3}, -1, -\frac{4}{3}, -24)$ *and* $(-\frac{2}{3}, -1, -\frac{4}{3}, 24)$.

Chapter 5

Error Analysis of Multilinear Terms using Linear Functions

Linearizations are commonly used to approximate monomial terms that are present within optimization problems. Pairwise products of bounded variables, for example, are handled within global optimization strategies by using McCormick [20] or RLT-type inequalities. For polynomial products of binary variables or continuous variables, a variety of linearization strategies exist ([16, 17, 18, 30, 3, 26, 19]). A common feature to all such approaches, both for continuous and binary variables, is the replacement of the monomial term with a continuous variable, and the subsequent employment of auxiliary constraints to relate the introduced variable to the represented product.

This effort differs from traditional approaches in that, it replaces a monomial term with a linear function in such a manner that no additional variables or constraints are used, and the associated error term is as small as possible. We, in fact, compute and verify the best such linear function and provide the worst case error. This is an approximation instead of relaxation, however this error compares favorably with those available from linearization methods. Specifically, for pairwise-products of bounded variables, we show that the error coincides with that of [2], while for the product of three or more binary variables, the error is theoretically superior to all linearizations. The error is also superior to the multilinear monomial linearization error computed in Section 4.

To elaborate, consider the multilinear monomial $m(\boldsymbol{x}) = \prod_{j=1}^{n} x_j$ defined on a box X' of (1.4). We seek to find $(\alpha_0, \boldsymbol{\alpha}) \in \mathbb{R}^{n+1}$, so that the linear function $L_{(\alpha_0,\boldsymbol{\alpha})} = \boldsymbol{\alpha}^T\boldsymbol{x} + \alpha_0$ minimizes

the maximum error $\left|\prod_{j=1}^{n} x_j - (\alpha^T \boldsymbol{x} + \alpha_0)\right|$ over $\boldsymbol{x} \in X'$. That is, we wish to solve:

$$\eta\left(X'\right) = \min_{(\alpha_0,\boldsymbol{\alpha})\in\mathbb{R}\times\mathbb{R}^n} \left\{ \max_{\boldsymbol{x}\in X'} \left|\prod_{j=1}^{n} x_j - (\boldsymbol{\alpha}^T \boldsymbol{x} + \alpha_0)\right| \right\}. \tag{5.1}$$

Observe that for the pairwise product of continuous variables, we have $n = 2$, while for the product of n binary variables, we have $L_j = 0$ and $U_j = 1$ for all j.

To date, we have proven, for the best linear function error:

1. Scaling result which, given an optimal solution to Problem (5.1), defines an optimal solution for any nonzero scaling of the x_j within X' of (1.4). Here for a vector $\boldsymbol{c} \in \mathbb{R}^n$ of nonzero scalars, X'_C of (4.4) is defined in terms of X' by component-wise multiplying the variables $\boldsymbol{x}$ of X' by $\boldsymbol{c}$.

   Lemma 5.0.1. *Given any $\boldsymbol{c} \in \mathbb{R}^n$ with $c_j \neq 0$ for all j, an $(\bar{\alpha}_0, \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{x}})$ is optimal to (5.1) with value $\eta\left(X'\right)$ if and only if $(\hat{\alpha}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{x}})$ is optimal to (5.2) with value $\eta\left(X'_C\right) = |\prod_{i=1}^{n} c_i|\, \eta\left(X'\right)$ and with $\hat{\alpha}_0 = (\prod_{i=1}^{n} c_i)\, \bar{\alpha}_0$, $\hat{\alpha}_j = \left(\prod_{i\neq j} c_i\right) \bar{\alpha}_j$ for all j, and $\hat{x}_j = c_j \bar{x}_j$ for all j.*

2. Solve Problem (5.1) for $n = 2$, including an explicit description of the unique function $L_{(\alpha_0,\boldsymbol{\alpha})} = \boldsymbol{\alpha}^T \boldsymbol{x} + \alpha_0$, and all $\boldsymbol{x} \in X'$ yielding this error.

   Theorem 5.0.2. *When $n = 2$, the error $\eta\left(X'\right)$ of (5.1) is $\eta\left([L_1, U_1] \times [L_2, U_2]\right) = \frac{(U_1 - L_1)(U_2 - L_2)}{4}$, and there exists a unique linear function affording this error, defined with $\alpha_0 = -\frac{(L_1+U_1)(L_2+U_2)}{4}$, $\alpha_1 = \left(\frac{L_2+U_2}{2}\right)$, and $\alpha_2 = \left(\frac{L_1+U_1}{2}\right)$, so that*

$$L_{(\alpha_0,\boldsymbol{\alpha})} = \left(\frac{L_2 + U_2}{2}\right) x_1 + \left(\frac{L_1 + U_1}{2}\right) x_2 - \frac{(L_1 + U_1)(L_2 + U_2)}{4}.$$

   Moreover, this maximum error is realized at only the four extreme points of X' given by $(x_1, x_2) = (L_1, L_2), (L_1, U_2), (U_1, L_2), (U_1, U_2)$.

3. Solve Problem (5.1) for $n \geq 2$ when $L_j = 0$ and $U_j = 1$ for each j, including an explicit description of the unique function $L_{(\alpha_0,\boldsymbol{\alpha})} = \boldsymbol{\alpha}^T \boldsymbol{x} + \alpha_0$, and all $\boldsymbol{x} \in X'$ yielding this error.

   Theorem 5.0.3. *When $L_j = 0$ and $U_j = 1$ for all j, the error $\eta\left(X'\right)$ of (5.1) is $\eta\left([0,1]^n\right) = \left(\frac{n-1}{2n}\right)$, and there exists a unique linear function affording this error, defined with $\alpha_0 = \left(\frac{1-n}{2n}\right)$*

and $\alpha_j = \left(\frac{1}{n}\right)$ *for all* j*, so that*

$$L_{(\alpha_0,\boldsymbol{\alpha})} = \sum_{j=1}^{n} \left(\frac{1}{n}\right) x_j + \left(\frac{1-n}{2n}\right).$$

Moreover, this maximum error is realized at only the $(n+2)$ *extreme points of* X' *given by* $\mathbf{0}$, $\mathbb{1}$, *and* $(\mathbb{1} - \boldsymbol{e}_j)$ *for all* j*, where* $\mathbf{0}$ *and* $\mathbb{1}$ *denote the vectors in* $\mathbb{R}^n$ *consisting of all value 0 and all value 1, respectively, and where* $\boldsymbol{e}_j$*, for each* j*, is the unit vector in* $\mathbb{R}^n$ *having a* 1 *in position* j *and* 0*'s elsewhere.*

This result is generalized, via Lemma 5.0.1, to solve Problem (5.1) for $n \geq 2$ when $L_j = 0$ or $U_j = 0$ for each j.

4. Solve Problem (5.1) for $n \geq 2$ when $L_j = 1$ and $U_j = r$ for each j and any $r > 1$, including an explicit description of the unique function $L_{(\alpha_0,\boldsymbol{\alpha})} = \boldsymbol{\alpha}^T\boldsymbol{x} + \alpha_0$, and all $\boldsymbol{x} \in X'$ yielding this error.

Theorem 5.0.4. *For any given* $r > 1$*, when* $L_j = 1$ *and* $U_j = r$ *for all* j*, the error* $\eta(X')$ *of* (5.1) *is* $\eta([1,r]^n) = \mathcal{F}_{r,n} \equiv \frac{1}{2}\left(1 + (r-1)qi^* - r^{i^*}\right)$*, and there exists a unique linear function affording this error, defined with* $\alpha_0 = -\frac{1}{2}\left(2nq + (r-1)qi^* - r^{i^*} - 1\right)$ *and* $\alpha_j = q$ *for all* j*, so that*

$$L_{(\alpha_0,\boldsymbol{\alpha})} = \sum_{j=1}^{n} qx_j - \frac{1}{2}\left(2nq + (r-1)qi^* - r^{i^*} - 1\right).$$

Here, $i^* = \lceil log_r(q) \rceil$ *with* $q = \left(\frac{r^n-1}{n(r-1)}\right)$*. Moreover, when* $log_r(q)$ *is not an integer, this maximum error is realized at only the* $\binom{n}{i^*} + 2$ *extreme points of* X' *given by* $\mathbb{1}$, $r\mathbb{1}$, *and the* $\binom{n}{i^*}$ *points containing exactly* i^* *entries of value* r *and* $(n - i^*)$ *entries of value* 1*, where* $\mathbb{1}$ *and* $r\mathbb{1}$ *denote the vectors in* $\mathbb{R}^n$ *consisting of all value* 1 *and all value* r*, respectively. However, when* $log_r(q)$ *is an integer, the maximum error is realized at* $\mathbb{1}$, $r\mathbb{1}$, *and every* $\boldsymbol{x} \in X'$ *such that at least* i^* *entries of* $\boldsymbol{x}$ *realize value* r *and at least* $(n - i^* - 1)$ *entries realize value 1.*

This result is generalized, via Lemma 5.0.1, to solve Problem (5.1) for $n \geq 2$ when each j has either $L_j > 0$ and $U_j = rL_j$ or has $U_j < 0$ and $L_j = rU_j$ for some $r > 1$.

5. Solve Problem (5.1) for $n \geq 2$ when $L_j = -1$ and $U_j = 1$ for each j, including an explicit description of the unique function $L_{(\alpha_0,\boldsymbol{\alpha})} = \boldsymbol{\alpha}^T\boldsymbol{x} + \alpha_0$, and all $\boldsymbol{x} \in X'$ yielding this error.

Theorem 5.0.5. *When $L_j = -1$ and $U_j = 1$ for all j, the error $\eta(X')$ of (5.1) is $\eta([-1,1]^n) = 1$, and there exists a unique linear function affording this error, defined with $\alpha_0 = 0$ and $\alpha_j = 0$ for all j, so that*

$$L_{(\alpha_0,\boldsymbol{\alpha})} = \sum_{j=1}^{n} 0x_j + 0.$$

Moreover, this maximum error is realized at only the 2^n extreme points of X'.

This result is generalized, via Lemma 5.0.1, to solve Problem (5.1) for $n \geq 2$ when $-L_j = U_j$ for each j.

We argue each of these results by firstly obtaining a tight lower bound on the error $\eta(X')$ of (5.1), then identifying the unique best linear function $L_{(\alpha_0,\boldsymbol{\alpha})}$ realizing this error in its worst case, and finally identifying all the points at which the best linear function attain its worst case error. In the second and last steps, we use the valid inequalities obtained, and the conditions that a nonnegative multilinear polynomial vanishes, from Section 2.1.

The structure of the monomial $\prod_{j=1}^{n} x_j$ and set X' in (1.4) allows the baseline cases of Table 5.1 to be extended. These extensions follow from Lemmas 4.1.1 and 5.0.1 found at the beginning of Section 4 and Chapter 5, respectively. Suppose that for some set of bounds L_j and U_j defining (1.4), we know conv(S), $\mu(X')$ of (4.1), and the set of points $(\boldsymbol{x}, y)$ at which $\mu(X')$ is realized. Then Lemma 2.1 allows us to compute conv(S), $\mu(X')$ of (4.1), and the set of points $(\boldsymbol{x}, y)$ at which $\mu(X')$ is realized when $\boldsymbol{x} \in X'$ of (1.4) is replaced with $\boldsymbol{x} \in X'_C$ of (4.4). Similarly, suppose that for some set of bounds L_j and U_j defining (1.4), we know $\eta(X')$ of (5.1) and the associated linear function(s) $L_{(\alpha_0,\boldsymbol{\alpha})}$, as well as the set of points $\boldsymbol{x} \in X'$ at which $\eta(X')$ is realized for each minimizing function. Then Lemma 3.1 allows us to compute this same information when $\boldsymbol{x} \in X'$ of (1.4) is replaced with $\boldsymbol{x} \in X'_C$ of (4.4). These more general results are presented as corollaries to the individual theorems.

5.1 Linear Function Replacement

Given a monomial term $\prod_{j=1}^{n} x_j$ with X' as in (1.4), we compute for various families of variable bounds $[L_j, U_j]$, the worst-case errors $\eta(X')$ of (5.1), the associated linear functions $L_{(\alpha_0,\boldsymbol{\alpha})} = \boldsymbol{\alpha}^T\boldsymbol{x} + \alpha_0$ affording these errors, and the collections of points $\boldsymbol{x} \in X'$ at which $\eta(X')$ is realized for each minimizing function. We begin with Lemma 5.0.1 that identifies the effects that

Table 5.1: Worst-Case Errors on Multilinear Term $\prod_{j=1}^{n} x_j$ over $\boldsymbol{x} \in X'$ of (1.4).

Size	Bounds	Convex Hull		Linear Function	
	$[L_j, U_j]$	Error $\mu(X')$ of (4.1)	Theorem	Error $\eta(X')$ of (5.1)	Theorem
$n = 2$	$[L_1, U_1], [L_2, U_2]$	$\frac{(U_1 - L_1)(U_2 - L_2)}{4}$	[2]	$\frac{(U_1 - L_1)(U_2 - L_2)}{4}$	5.0.2
$n \geq 2$	$[0, 1]$	$(n-1)n^{\frac{n}{1-n}}$	4.0.1	$\frac{n-1}{2n}$	5.0.3
$n \geq 2$	$[1, r]$	$\max\{\mathcal{D}_{r,n}, \mathcal{E}_{r,n}\}$	4.0.6	$\mathcal{F}_{r,n}$	5.0.4
$n \geq 2$	$[-1, 1]$	$1 + \left(\frac{n-2}{n}\right)^n$	4.0.7	1	5.0.5

scaling the variables $\boldsymbol{x} \in X'$ of (1.4) to obtain $\boldsymbol{x} \in X'_C$ of (4.4) has upon these results. In particular, suppose that for some set X' of (1.4), a "best" linear function $\bar{\boldsymbol{\alpha}}^T \boldsymbol{x} + \bar{\alpha}_0$ has been computed to obtain the error $\eta(X')$ of (5.1). Then, for any vector $\boldsymbol{c} \in \mathbb{R}^n$ of nonzero scalars, we can readily compute a linear function, say $\hat{\boldsymbol{\alpha}}^T \boldsymbol{x} + \hat{\alpha}_0$, which minimizes over $(\alpha_0, \boldsymbol{\alpha}) \in \mathbb{R} \times \mathbb{R}^n$, the maximum error $\left| \prod_{j=1}^{n} x_j - \left(\boldsymbol{\alpha}^T \boldsymbol{x} + \alpha_0 \right) \right|$ over $\boldsymbol{x} \in X'_C$ of (4.4), expressed as

$$\eta(X'_C) = \min_{(\alpha_0, \boldsymbol{\alpha}) \in \mathbb{R} \times \mathbb{R}^n} \left\{ \max_{\boldsymbol{x} \in X'_C} \left| \prod_{j=1}^{n} x_j - (\boldsymbol{\alpha}^T \boldsymbol{x} + \alpha_0) \right| \right\}. \tag{5.2}$$

Using the vector $\boldsymbol{c}$, such a linear function is available in terms of $\bar{\boldsymbol{\alpha}}^T \boldsymbol{x} + \bar{\alpha}_0$. In addition, the error $\eta(X'_C)$ of (5.2) is available in terms of $\eta(X')$ of (5.1), and the collection of points $\boldsymbol{x} \in X'_C$ at which $\eta(X'_C)$ is realized is available in terms of the computed points of $\boldsymbol{x} \in X'$ affording $\eta(X')$.

Proof of Lemma 5.0.1. Since a point $(\bar{\alpha}_0, \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{x}})$ is feasible to (5.1) if and only if $(\hat{\alpha}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{x}})$ defined in terms of $(\bar{\alpha}_0, \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{x}})$ as in the Lemma is feasible to (5.2), the proof is to show that the points $(\bar{\alpha}_0, \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{x}})$ and $(\hat{\alpha}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{x}})$ have $\left| \prod_{i=1}^{n} c_i \right| \left| \prod_{j=1}^{n} \bar{x}_j - (\bar{\boldsymbol{\alpha}}^T \bar{\boldsymbol{x}} + \bar{\alpha}_0) \right| = \left| \prod_{j=1}^{n} \hat{x}_j - (\hat{\boldsymbol{\alpha}}^T \hat{\boldsymbol{x}} + \hat{\alpha}_0) \right|$. We have

$$\begin{aligned} \left| \prod_{i=1}^{n} c_i \right| \left| \prod_{j=1}^{n} \bar{x}_j - (\bar{\boldsymbol{\alpha}}^T \bar{\boldsymbol{x}} + \bar{\alpha}_0) \right| &= \left| \prod_{j=1}^{n} (c_j \bar{x}_j) - \left[\sum_{j=1}^{n} \left(\prod_{i=1}^{n} c_i \right) \bar{\alpha}_j \bar{x}_j + \left(\prod_{i=1}^{n} c_i \right) \bar{\alpha}_0 \right] \right| \\ &= \left| \prod_{j=1}^{n} \hat{x}_j - \left(\sum_{j=1}^{n} \hat{\alpha}_j \hat{x}_j + \hat{\alpha}_0 \right) \right|, \end{aligned}$$

where the first equality distributes the expression $|\prod_{i=1}^{n} c_i|$, and the second equality follows from the definition of $(\hat{\alpha}_0, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{x}})$ in terms of $(\bar{\alpha}_0, \bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{x}})$. The proof is complete. □

Lemma 5.0.1 allows us to focus on specific bounds $[L_j, U_j]$ within (1.4) to compute $\eta(X')$ of (5.1), and to then generalize these values via scalings as in (4.4) to compute $\eta(X'_C)$ of (5.2). These scalings apply to Theorems 5.0.3, 5.0.5, and 5.0.4.

We begin with the case of (5.1) having $n = 2$ in Theorem 5.0.2 below.

Proof of Theorem 5.0.2. The proof consists of three parts: the first part establishes $\frac{(U_1-L_1)(U_2-L_2)}{4}$ as a lower bound on $\eta(X')$, the second part shows that the only linear function $L_{(\alpha_0,\boldsymbol{\alpha})}$ that can possibly yield this bound is defined in terms of the stated $(\alpha_0, \alpha_1, \alpha_2)$, and the third part shows that this bound is indeed realized for the given $L_{(\alpha_0,\boldsymbol{\alpha})}$ at only the four extreme points of $X' = [L_1, U_1] \times [L_2, U_2]$.

- When $n = 2$, the error $\eta(X')$ of (5.1) is given by

$$\eta(X') = \min_{(\alpha_0,\alpha_1,\alpha_2)} \left\{ \max_{\boldsymbol{x}\in X'} \{|x_1x_2 - (\alpha_1x_1 + \alpha_2x_2 + \alpha_0)|\} \right\}. \tag{5.3}$$

  Regardless of the values of $(\alpha_0, \alpha_1, \alpha_2)$, the four extreme points of X' give us that

$$\begin{aligned}
&4\left\{ \max_{\boldsymbol{x}\in X'} \{|x_1x_2 - (\alpha_1x_1 + \alpha_2x_2 + \alpha_0)|\} \right\} \\
&\geq |L_1L_2 - (\alpha_1L_1 + \alpha_2L_2 + \alpha_0)| + |L_1U_2 - (\alpha_1L_1 + \alpha_2U_2 + \alpha_0)| \\
&\quad + |U_1L_2 - (\alpha_1U_1 + \alpha_2L_2 + \alpha_0)| + |U_1U_2 - (\alpha_1U_1 + \alpha_2U_2 + \alpha_0)| \\
&= |L_1L_2 - (\alpha_1L_1 + \alpha_2L_2 + \alpha_0)| + |-L_1U_2 + (\alpha_1L_1 + \alpha_2U_2 + \alpha_0)| \\
&\quad + |-U_1L_2 + (\alpha_1U_1 + \alpha_2L_2 + \alpha_0)| + |U_1U_2 - (\alpha_1U_1 + \alpha_2U_2 + \alpha_0)| \\
&\geq [L_1L_2 - (\alpha_1L_1 + \alpha_2L_2 + \alpha_0)] + [-L_1U_2 + (\alpha_1L_1 + \alpha_2U_2 + \alpha_0)] \\
&\quad + [-U_1L_2 + (\alpha_1U_1 + \alpha_2L_2 + \alpha_0)] + [U_1U_2 - (\alpha_1U_1 + \alpha_2U_2 + \alpha_0)] \\
&= (U_1 - L_1)(U_2 - L_2).
\end{aligned} \tag{5.4}$$

  Then (5.3) has $4\eta(X') \geq (U_1 - L_1)(U_2 - L_2)$ so that $\eta(X') \geq \frac{(U_1-L_1)(U_2-L_2)}{4}$.

- Suppose there exists an $(\alpha_0, \alpha_1, \alpha_2)$ so that $\eta(X')$ from (5.3) has $\eta(X') = \frac{(U_1-L_1)(U_2-L_2)}{4}$. Then both inequalities of (5.4) must be satisfied at equality for this $(\alpha_0, \alpha_1, \alpha_2)$. Since each absolute value expression in the first inequality would be bounded above by $\eta(X')$, this inequality would be satisfied at equality if and only if each of these four expressions is equal to

$\frac{(U_1-L_1)(U_2-L_2)}{4}$. But then the second inequality would also be satisfied at equality if and only if each of the following four equations holds:

$$L_1L_2 - (\alpha_1L_1 + \alpha_2L_2 + \alpha_0) = \eta(X'), \quad -L_1U_2 + (\alpha_1L_1 + \alpha_2U_2 + \alpha_0) = \eta(X')$$
$$-U_1L_2 + (\alpha_1U_1 + \alpha_2L_2 + \alpha_0) = \eta(X'), \quad U_1U_2 - (\alpha_1U_1 + \alpha_2U_2 + \alpha_0) = \eta(X'),$$

with $\eta(X') = \frac{(U_1-L_1)(U_2-L_2)}{4}$. This system of four linear equations in three unknowns has the unique solution $\alpha_0 = -\frac{(L_1+U_1)(L_2+U_2)}{4}$, $\alpha_1 = \left(\frac{L_2+U_2}{2}\right)$, $\alpha_2 = \left(\frac{L_1+U_1}{2}\right)$, and $\eta(X') = \frac{(U_1-L_1)(U_2-L_2)}{4}$, as stated in the Theorem.

- Let $(\alpha_0, \alpha_1, \alpha_2)$ be as stated in the Theorem and derived above. Then the inner maximization problem of (5.3) is
$$\max_{\boldsymbol{x}\in X'} \left\{ \left| \left(x_1 - \frac{L_1+U_1}{2}\right)\left(x_2 - \frac{L_2+U_2}{2}\right) \right| \right\},$$
and has maximum value of $\frac{(U_1-L_1)(U_2-L_2)}{4}$ at only the four extreme points of X'.

□

Observe an interesting relationship between the error associated with approximating the monomial term $\prod_{j=1}^n x_j$ for $n = 2$ when using the convex hull representation of [20, 30] as in (4.5), as opposed to that realized when using the linear function $L_{(\alpha_0,\boldsymbol{\alpha})}$ as provided in Theorem 5.0.2. Both approximations have a worse-case error of $\frac{(U_1-L_1)(U_2-L_2)}{4}$, but this error occurs in the convex hull representation at the two points $(x_1, x_2, y) = (\frac{L_1+U_1}{2}, \frac{L_2+U_2}{2}, \frac{U_1L_2+L_1U_2}{2})$ and $(x_1, x_2, y) = (\frac{L_1+U_1}{2}, \frac{L_2+U_2}{2}, \frac{L_1L_2+U_1U_2}{2})$, while it occurs for the linear function at the four points $(x_1, x_2) = (L_1, L_2)$, $(x_1, x_2) = (L_1, U_2)$, $(x_1, x_2) = (U_1, L_2)$, and $(x_1, x_2) = (U_1, U_2)$. The representation (4.5) is exact so that $y = x_1x_2$ at each of the four points where the linear function realizes its maximum error and, conversely, the linear function is exact, taking value $\frac{(L_1+U_1)(L_2+U_2)}{4}$ when $(x_1, x_2) = \left(\frac{L_1+U_1}{2}, \frac{L_2+U_2}{2}\right)$, where the representation [20, 30] realizes its largest error.

Example 5. *Consider* $\prod_{j=1}^n x_j$ *in* $n = 2$ *variables* x_1 *and* x_2 *when* $L_1 = 2, L_2 = 3, U_1 = 5, U_2 = 7$ *in* (1.4). *Then the set conv*(S) *of* (4.5) *is below.*

$$conv(S) = \left\{ \begin{array}{ll} (x_1, x_2, y) \in \mathbb{R}^3 : & 7x_1 + 5x_2 - 35 \le y \le 3x_1 + 5x_2 - 15 \\ & 3x_1 + 2x_2 - 6 \le y \le 7x_1 + 2x_2 - 14 \end{array} \right\}$$

Reference [2] gives the error $\mu(X')$ of (4.1) to be $\mu(X') = 3$, with this error occurring at $(x_1, x_2, y) = (\frac{7}{2}, 5, \frac{29}{2})$ and $(x_1, x_2, y) = (\frac{7}{2}, 5, \frac{41}{2})$. Theorem 5.0.2 gives the error $\eta(X')$ of (5.1) to be $\eta(X') = 3$, and to be associated with the linear function $L_{(\alpha_0,\boldsymbol{\alpha})} = 5x_1 + \frac{7}{2}x_2 - \frac{35}{2}$. In addition, Theorem 5.0.2 states that this error occurs at only the four points $(x_1, x_2) = (2,3), (2,7), (5,3)$, and $(5,7)$. The convex hull approximation is exact at these four points, while the linear approximation is exact at $(x_1, x_2) = (\frac{7}{2}, 5)$.

5.2 Proof of Theorem 5.0.3, where $X' = X = [0,1]^n$

Theorem 5.0.3 addresses the error $\eta(X')$ of (5.1) when $L_j = 0$ and $U_j = 1$ for all j in (1.4). The proof follows a similar structure to that of Theorem 5.0.2. When $n = 2$, and $L_1 = L_2 = 0$ and $U_1 = U_2 = 1$ in Theorem 5.0.2, we have that Theorems 5.0.2 and 5.0.3 coincide.

Proof of Theorem 5.0.3. Following the proof of Theorem 5.0.2, the argument consists of three parts: the first part establishes $\left(\frac{n-1}{2n}\right)$ as a lower bound on $\eta(X')$, the second part shows that the only linear function $L_{(\alpha_0,\boldsymbol{\alpha})}$ that can possibly yield this bound is defined in terms of the stated $(\alpha_0, \boldsymbol{\alpha})$, and the third part shows that this bound is indeed realized for the given $L_{(\alpha_0,\boldsymbol{\alpha})}$ at only the stated $(n+2)$ extreme points of $X' = [0,1]^n$.

- Regardless of the values of $(\alpha_0, \boldsymbol{\alpha})$, the stated $(n+2)$ extreme points of X', with multiplier $(n-1)$ on extreme point $\mathbb{1}$ and multiplier 1 on all other points, give us that

$$
\begin{aligned}
&(2n)\left\{\max_{\boldsymbol{x}\in X'}\left|\prod_{j=1}^{n} x_j - (\boldsymbol{\alpha}^T\boldsymbol{x} + \alpha_0)\right|\right\} \\
&\quad\geq |-\alpha_0| + (n-1)\left|1 - \left(\sum_{j=1}^{n}\alpha_j + \alpha_0\right)\right| + \sum_{j=1}^{n}\left|0 - \left(\sum_{k\neq j}\alpha_k + \alpha_0\right)\right| \\
&\quad= |-\alpha_0| + (n-1)\left|1 - \sum_{j=1}^{n}\alpha_j - \alpha_0\right| + \sum_{j=1}^{n}\left|\sum_{k\neq j}\alpha_k + \alpha_0\right| \\
&\quad\geq -\alpha_0 + (n-1)\left[1 - \sum_{j=1}^{n}\alpha_j - \alpha_0\right] + \sum_{j=1}^{n}\left[\sum_{k\neq j}\alpha_k + \alpha_0\right] \\
&\quad= n-1. \qquad (5.5)
\end{aligned}
$$

Then (5.1) has $(2n)\eta(X') \geq n-1$ so that $\eta(X') \geq \left(\frac{n-1}{2n}\right)$.

- Suppose there exists an $(\alpha_0, \boldsymbol{\alpha})$ so that $\eta(X')$ from (5.1) has $\eta(X') = \left(\frac{n-1}{2n}\right)$. Then both inequalities of (5.5) must be satisfied at equality for this $(\alpha_0, \boldsymbol{\alpha})$. Since each absolute value expression in the first inequality would be bounded above by $\eta(X')$, this inequality would be satisfied at equality if and only if each of these $(n+2)$ expressions is equal to $\left(\frac{n-1}{2n}\right)$. But then the second inequality would also be satisfied at equality if and only if each of the following $(n+2)$ equations holds:

$$-\alpha_0 = \left(\frac{n-1}{2n}\right), \;\; 1 - \sum_{j=1}^{n} \alpha_j - \alpha_0 = \left(\frac{n-1}{2n}\right), \;\; \sum_{k \neq j} \alpha_k + \alpha_0 = \left(\frac{n-1}{2n}\right) \;\; \forall \; j.$$

  We have $\alpha_0 = \left(\frac{1-n}{2n}\right)$ by the first equation. For each j, add the equation of the third set to the second equation to obtain $\alpha_j = 1 - 2\left(\frac{n-1}{2n}\right) = \left(\frac{1}{n}\right)$. Thus, this system of $(n+2)$ linear equations in $(n+1)$ unknowns has the unique solution $\alpha_0 = \left(\frac{1-n}{2n}\right)$ and $\alpha_j = \left(\frac{1}{n}\right)$ for all j, as stated in the Theorem.

- Let $(\alpha_0, \boldsymbol{\alpha})$ be as stated in the Theorem and derived above. Then the inner maximization problem of (5.1) is

$$\max_{\boldsymbol{x} \in X'} \left\{ \left| \prod_{j=1}^{n} x_j - \sum_{j=1}^{n} \left(\frac{1}{n}\right) x_j + \left(\frac{n-1}{2n}\right) \right| \right\}. \tag{5.6}$$

  We have

$$\frac{1-n}{n} \leq \prod_{j=1}^{n} x_j - \sum_{j=1}^{n} \left(\frac{1}{n}\right) x_j \leq 0. \tag{5.7}$$

  The left inequality is computed as a surrogate of the two inequalities $\left[0 \leq \prod_{j=1}^{n} x_j\right]$ and $\left[\sum_{j=1}^{n} x_j - (n-1) \leq \prod_{j=1}^{n} x_j\right]$ from (4.6) using multipliers $\left(\frac{n-1}{n}\right)$ and $\left(\frac{1}{n}\right)$, respectively, with $y = \prod_{j=1}^{n} x_j$. The right inequality is a surrogate of the n inequalities $\left[\prod_{j=1}^{n} x_j \leq x_j\right]$ for all j from (4.6) using multipliers $\left(\frac{1}{n}\right)$, again with $y' = \prod_{j=1}^{n} x_j$. Adding $\left(\frac{n-1}{2n}\right)$ to each expression of (5.7), we obtain that the absolute value expression of (5.6) is bounded above by $\left(\frac{n-1}{2n}\right)$. Since the two inequalities in (4.6) used to compute the left inequality of (5.7) are both satisfied with equality at precisely the n points of X' given by $(1 - \boldsymbol{e}_j)$ for all j, and since the n inequalities in (4.6) used to compute the right inequality of (5.7) are all satisfied with equality at precisely the 2 points of X' given by $\mathbf{0}$ and $\mathbb{1}$, this maximum value of $\left(\frac{n-1}{2n}\right)$ is realized in (5.6) at precisely these $(n+2)$ extreme points of X'.

The proof is complete. □

Lemma 5.0.1 allows a generalization of Theorem 5.0.3 for nonzero scalings $\boldsymbol{c}$ of the variables $\boldsymbol{x}$. This generalization can be interpreted as stating that, provided $L_jU_j = 0$ for all j, an optimal solution to (5.1) is available in terms of the solution given in Theorem 5.0.3. The result follows by letting $c_j = L_j$ if $U_j = 0$ and $c_j = U_j$ if $L_j = 0$ so that $c_j = (L_j + U_j)$ for all j. The formal statement is given below, without proof, as a corollary to Theorem 5.0.3.

Corollary 5.2.1. *When $L_jU_j = 0$ for all j, the error $\eta(X'_C)$ of (5.2) is $\eta(X'_C) = \left|\prod_{j=1}^n c_j\right| \left(\frac{n-1}{2n}\right)$, and there exists a unique linear function affording this error, defined with $\alpha_0 = \left(\prod_{j=1}^n c_j\right)\left(\frac{1-n}{2n}\right)$ and $\alpha_j = \left(\prod_{i\neq j} c_i\right)\left(\frac{1}{n}\right)$ for all j, so that*

$$L_{(\alpha_0,\boldsymbol{\alpha})} = \sum_{j=1}^n \left(\prod_{i\neq j} c_i\right)\left(\frac{1}{n}\right) x_j + \left(\prod_{j=1}^n c_j\right)\left(\frac{1-n}{2n}\right).$$

Here, $\boldsymbol{c}$ has $c_j = (L_j + U_j)$ for all j. Moreover, this worst-case error is realized at only the $(n+2)$ extreme points of X'_C given by $\mathbf{0}, \boldsymbol{c}$, and $(\boldsymbol{c} - c_j\boldsymbol{e}_j)$ for each j.

We can now compare the errors associated with approximating $\prod_{j=1}^n x_j$ having $L_jU_j = 0$ for all j in (1.4) using the approach of the convex hull representation as in (4.8) with that of the linear approximation $L_{(\alpha_0,\boldsymbol{\alpha})}$ as in Corollary 5.2.1. Theorem 4.0.1 and scaling Lemma 4.1.1 give the first error as $\mu(X') = \left|\prod_{j=1}^n c_j\right|(n-1)n^{\frac{n}{1-n}}$, while Corollary 5.2.1 gives the second error as $\eta(X'_C) = \left|\prod_{j=1}^n c_j\right|\left(\frac{n-1}{2n}\right)$, with $c_j = (L_j + U_j)$ for all j. These errors coincide when $n = 2$, but have $\mu(X') > \eta(X'_C)$ for $n \geq 3$. Asymptotically, we have $\lim_{n\to\infty}\left(\left|\prod_{j=1}^n c_j\right|(n-1)n^{\frac{n}{1-n}}\right) = \left|\prod_{j=1}^n c_j\right|$ while $\lim_{n\to\infty}\left(\left|\prod_{j=1}^n c_j\right|\left(\frac{n-1}{2n}\right)\right) = \frac{\left|\prod_{j=1}^n c_j\right|}{2}$.

The below example considers an instance of $\prod_{j=1}^n x_j$ where all variables x_j in (1.4) have $L_j = 0$. Then $I^1 = \{1,\ldots,n\}$ and $I^2 = \emptyset$ in (4.8).

Example 6. *Consider the monomial $\prod_{j=1}^n x_j$ in $n = 3$ variables x_j when $L_1 = L_2 = L_3 = 0$, and $U_1 = 2$, $U_2 = 3$, $U_3 = 6$ in (1.4). Then the set (4.8) with $I^1 = \{1,2,3\}$ and $I^2 = \emptyset$ is below.*

$$\left\{ \begin{array}{ll} (x_1,x_2,x_3,y') \in \mathbb{R}^4 : & \\ & y' \leq 18x_1, y' \leq 12x_2, y' \leq 6x_3 \\ & x_1 \leq 2, x_2 \leq 3, x_3 \leq 6 \\ & y' \geq 18x_1 + 12x_2 + 6x_3 - 72 \\ & y' \geq 0 \end{array} \right\}$$

Theorem 4.0.1 gives the error $\mu(X')$ *of* (4.1) *as* $\mu(X') = 8\sqrt{3}$, *and that this error occurs at* $(x_1, x_2, x_3, y') = (\frac{2\sqrt{3}}{3}, \sqrt{3}, 2\sqrt{3}, 12\sqrt{3})$. *Corollary 5.2.1 with* $(c_1, c_2, c_3) = (2, 3, 6)$ *gives the error* $\eta(X'_C)$ *of* (5.2) *as* $\eta(X'_C) = 12$, *which is associated with the linear function* $L_{(\alpha_0, \boldsymbol{\alpha})} = 6x_1 + 4x_2 + 2x_3 - 12$. *In addition, Corollary 5.2.1 states that this error occurs at only the five points* $(x_1, x_2, x_3) = (0, 0, 0), (0, 3, 6), (2, 0, 6), (2, 3, 0),$ *and* $(2, 3, 6)$.

Example 7 below considers an instance of $\prod_{j=1}^{n} x_j$ with both nonnegative and nonpositive variables x_j in (1.4). Then neither of the sets I^1 nor I^2 in (4.8) is empty.

Example 7. *Consider the monomial* $\prod_{j=1}^{n} x_j$ *in* $n = 3$ *variables when* $L_1 = U_2 = L_3 = 0$, *and* $U_1 = 4$, $L_2 = -3$, $U_3 = 6$ *in* (1.4). *Then the set* (4.8) *with* $I^1 = \{1, 3\}$ *and* $I^2 = \{2\}$ *is below.*

$$\left\{ \begin{array}{ll} (x_1, x_2, x_3, y') \in \mathbb{R}^4 : & \\ & -y' \leq 18x_1, -y' \leq -24x_2, -y' \leq 12x_3 \\ & x_1 \leq 4, x_2 \geq -3, x_3 \leq 6 \\ & -y' \geq 18x_1 - 24x_2 + 12x_3 - 144 \\ & -y' \geq 0 \end{array} \right\}$$

Theorem 4.0.1 gives the error $\mu(X')$ *of* (4.1) *as* $\mu(X') = 16\sqrt{3}$, *and that this error occurs at* $(x_1, x_2, x_3, y') = (\frac{4\sqrt{3}}{3}, -\sqrt{3}, 2\sqrt{3}, -24\sqrt{3})$. *Corollary 5.2.1 with* $(c_1, c_2, c_3) = (4, -3, 6)$ *gives the error* $\eta(X'_C)$ *of* (5.2) *as* $\eta(X'_C) = 24$, *which is associated with the linear function* $L_{(\alpha_0, \boldsymbol{\alpha})} = -6x_1 + 8x_2 - 4x_3 + 24$. *In addition, Corollary 5.2.1 states that this error occurs at only the five points* $(x_1, x_2, x_3) = (0, 0, 0), (0, -3, 6), (4, 0, 6), (4, -3, 0),$ *and* $(4, -3, 6)$.

5.3 Proof of Theorem 5.0.4, where $X' = X = [1, r]^n$

Given any scalar $r > 1$, Theorem 5.0.4 below identifies the error $\eta(X')$ of (5.1) when $L_j = 1$ and $U_j = r$ for all j.

Proof of Theorem 5.0.4. Following the proofs of Theorems 5.0.2, 5.0.3, and 5.0.5, the argument consists of three parts: the first part establishes $\frac{1}{2}\left(1 + (r-1)qi^* - r^{i^*}\right)$ as a lower bound on $\eta(X')$, the second part shows that the only linear function $L_{(\alpha_0, \boldsymbol{\alpha})}$ that can possibly yield this bound is defined in terms of the stated $(\alpha_0, \boldsymbol{\alpha})$, and the third part shows that this bound is indeed realized for the given $L_{(\alpha_0, \boldsymbol{\alpha})}$ at only the stated points of $X' = [1, r]^n$.

- For any $p \in \{1, \ldots, n-1\}$, consider the $\binom{n}{p} + 2$ extreme points of X' given by $\mathbb{1}$, $r\mathbb{1}$, and the $\binom{n}{p}$ vectors in $\mathbb{R}^n$ containing exactly p entries of value r and $(n-p)$ entries of value 1, with $\mathbb{1}$ and $r\mathbb{1}$ as defined in the Theorem. Let $\boldsymbol{x}^j$ for $j = 1, \ldots, \binom{n}{p}$ denote the latter $\binom{n}{p}$ extreme points, with x_k^j representing entry k of $\boldsymbol{x}^j$. Regardless of the values of $(\alpha_0, \boldsymbol{\alpha})$, these $\binom{n}{p} + 2$ extreme points of X', with multiplier $\binom{n-1}{p}$ on extreme point $\mathbb{1}$, multiplier $\binom{n-1}{p-1}$ on extreme point $r\mathbb{1}$, and multiplier 1 on all remaining points, give us that

$$
\begin{aligned}
&2\binom{n}{p}\left\{\max_{\boldsymbol{x}\in X'}\left|\prod_{j=1}^{n} x_j - (\boldsymbol{\alpha}^T\boldsymbol{x} + \alpha_0)\right|\right\} \\
&\quad\geq \binom{n-1}{p}\left|1 - \left(\sum_{k=1}^{n}\alpha_k + \alpha_0\right)\right| + \binom{n-1}{p-1}\left|r^n - \left(\sum_{k=1}^{n}\alpha_k r + \alpha_0\right)\right| \\
&\qquad + \sum_{j=1}^{\binom{n}{p}}\left|r^p - \left(\sum_{k=1}^{n}\alpha_k x_k^j + \alpha_0\right)\right| \\
&\quad= \binom{n-1}{p}\left|1 - \sum_{k=1}^{n}\alpha_k - \alpha_0\right| + \binom{n-1}{p-1}\left|r^n - \sum_{k=1}^{n}\alpha_k r - \alpha_0\right| \\
&\qquad + \sum_{j=1}^{\binom{n}{p}}\left|-r^p + \sum_{k=1}^{n}\alpha_k x_k^j + \alpha_0\right| \\
&\quad\geq \binom{n-1}{p}\left[1 - \sum_{k=1}^{n}\alpha_k - \alpha_0\right] + \binom{n-1}{p-1}\left[r^n - \sum_{k=1}^{n}\alpha_k r - \alpha_0\right] \\
&\qquad + \sum_{j=1}^{\binom{n}{p}}\left[-r^p + \sum_{k=1}^{n}\alpha_k x_k^j + \alpha_0\right] \\
&\quad= \left[\binom{n-1}{p} + \binom{n-1}{p-1}r^n - \binom{n}{p}r^p\right] + \left[-\binom{n-1}{p} - \binom{n-1}{p-1} + \binom{n}{p}\right]\alpha_0 \\
&\qquad + \sum_{k=1}^{n}\left[-\binom{n-1}{p} - \binom{n-1}{p-1}r + \binom{n}{p}\left(\frac{p}{n}\right)r + \binom{n}{p}\left(\frac{n-p}{n}\right)\right]\alpha_k \\
&\quad= \binom{n-1}{p} + \binom{n-1}{p-1}r^n - \binom{n}{p}r^p. \qquad (5.8)
\end{aligned}
$$

Then (5.1) has $2\binom{n}{p}\eta(X') \geq \binom{n-1}{p} + \binom{n-1}{p-1}r^n - \binom{n}{p}r^p$ so that $\eta(X') \geq \frac{1}{2}\left(\frac{n-p}{n} + \frac{p}{n}r^n - r^p\right) = \frac{1}{2}\left(1 + (r-1)qp - r^p\right)$. Since $1 < q < r^{n-1}$, then $1 \leq i^* \leq n-1$ and we can let $p = i^*$ to obtain $\eta(X') \geq \frac{1}{2}\left[(1 + (r-1)qi^* - r^{i^*}\right]$.

- Suppose there exists an $(\alpha_0, \boldsymbol{\alpha})$ so that $\eta(X')$ from (5.1) has $\eta(X') = \frac{1}{2}\left[(1 + (r-1)qi^* - r^{i^*}\right]$. Then both inequalities of (5.8) with $p = i^*$ must be satisfied at equality for this $(\alpha_0, \boldsymbol{\alpha})$. Since

each absolute value expression in the first inequality would be bounded above by $\eta(X')$, this inequality would be satisfied at equality if and only if each of these $\binom{n}{p}+2$ expressions is equal to $\frac{1}{2}\left[(1+(r-1)qi^*-r^{i^*}\right]$. But then the second inequality would also be satisfied at equality if and only if each of the following $\binom{n}{p}+2$ equations holds:

$$1-\sum_{k=1}^{n}\alpha_k-\alpha_0=\eta(X'),\ r^n-\sum_{k=1}^{n}\alpha_k r-\alpha_0=\eta(X'),$$
$$\text{and } -r^p+\sum_{k=1}^{n}\alpha_k x_k^j+\alpha_0=\eta(X')\ \forall\ j=1,\ldots,\binom{n}{p}, \tag{5.9}$$

with $\eta(X')=\frac{1}{2}\left[(1+(r-1)qi^*-r^{i^*}\right]$. Subtract r times the first equation of (5.9) from the second equation to obtain $r^n-r+(r-1)\alpha_0=(1-r)\eta(X')$ so that, upon substituting $\eta(X')=\frac{1}{2}\left[(1+(r-1)qi^*-r^{i^*}\right]$ and recalling that $q=\left(\frac{r^n-1}{n(r-1)}\right)$, we obtain $\alpha_0=-\frac{1}{2}\left(2nq+(r-1)qi^*-r^{i^*}-1\right)$. Next, for any distinct $s,t\in\{1,\ldots,n\}$, consider any extreme point of X' containing exactly p entries of value r and $(n-p)$ entries of value 1, say $\boldsymbol{x}^{j_1}$, for which $x_s^{j_1}=r$ and $x_t^{j_1}=1$, and define extreme point $\boldsymbol{x}^{j_2}$ from $\boldsymbol{x}^{j_1}$ by interchanging entries s and t so that $x_s^{j_2}=1$ and $x_t^{j_2}=r$. From the last family of $\binom{n}{p}$ equations of (5.9), subtract that equation associated with $\boldsymbol{x}^{j_1}$ from that equation associated with $\boldsymbol{x}^{j_2}$ to obtain $(1-r)\alpha_s+(r-1)\alpha_t=0$ so that $\alpha_s=\alpha_t$. Since s and t were arbitrarily selected, we have $\alpha_s=\alpha_t$ for all distinct $s,t\in\{1,\ldots,n\}$. Then the first equation of (5.9) gives $1-n\alpha_j=\eta(X')+\alpha_0=1-nq$ for all j so that $\alpha_j=q$ for all j. Thus, this system of $\binom{n}{p}+2$ linear equations in $(n+1)$ unknowns has the unique solution for $(\alpha_0,\boldsymbol{\alpha})$ as stated in the Theorem.

- Let $(\alpha_0,\boldsymbol{\alpha})$, i^*, and q be as stated in the Theorem and derived above. Then the inner maximization problem of (5.1) is

$$\max_{\boldsymbol{x}\in X'}\left\{\left|\prod_{j=1}^{n}x_j-\sum_{j=1}^{n}qx_j+\frac{1}{2}\left(2nq+(r-1)qi^*-r^{i^*}-1\right)\right|\right\}. \tag{5.10}$$

We have

$$-\left(1+(r-1)qi^*-r^{i^*}\right)\le\prod_{j=1}^{n}x_j-\sum_{j=1}^{n}qx_j+nq-1\le 0. \tag{5.11}$$

(i) The left inequality of (5.11) follows from the inequalities

$$r^i \left(\sum_{j=1}^{n} x_j - (n-1) - i(r-1) \right) \leq \prod_{j=1}^{n} x_j \; \forall \; i = 0, \ldots, n-1 \tag{5.12}$$

of (4.10), which hold for all $\boldsymbol{x} \in X'$, for the following reasons. When $\log_r(q)$ is not an integer, inequalities (5.12) for $i = (i^* - 1)$ and $i = i^*$ are

$$r^{(i^*-1)} \left(\sum_{j=1}^{n} x_j - (n-1) - (i^*-1)(r-1) \right) \leq \prod_{j=1}^{n} x_j \tag{5.13}$$

and

$$r^{i^*} \left(\sum_{j=1}^{n} x_j - (n-1) - i^*(r-1) \right) \leq \prod_{j=1}^{n} x_j, \tag{5.14}$$

respectively. The left inequality of (5.11) is then a surrogate of (5.13) and (5.14) using nonnegative multipliers $\left(\frac{r^{i^*} - q}{r^{i^*} - r^{(i^*-1)}} \right)$ and $\left(\frac{q - r^{(i^*-1)}}{r^{i^*} - r^{(i^*-1)}} \right)$, respectively. When $\log_r(q)$ is an integer so that $i^* = \log_r(q)$, then inequality (5.12) for $i = i^* = \log_r(q)$ takes the form

$$q \left(\sum_{j=1}^{n} x_j - (n-1) - i^*(r-1) \right) \leq \prod_{j=1}^{n} x_j, \tag{5.15}$$

which is the left inequality of (5.11).

(ii) The right inequality of (5.11) also follows from inequalities of (4.9):

$$\prod_{j=1}^{n} x_j \leq \sum_{j=1}^{n} r^{j-1} x_{\sigma(j)} - nq + 1 \; \forall n\text{-permutation } \sigma \tag{5.16}$$

hold for all $\boldsymbol{x} \in X'$. Since each x_j appears exactly $(n-1)!$ times with coefficient r^k for $k \in \{0, \ldots, n-1\}$ within (5.16), a surrogate of the equations of (5.16) using multipliers $\left(\frac{1}{n!} \right)$ gives the right inequality of (5.11).

Adding $\eta(X') = \frac{1}{2}\left(1 + (r-1)qi^* - r^{i^*}\right)$ to each expression of (5.11), we obtain that the absolute value expression of (5.10) is bounded above by $\eta(X')$. Now, observe that for each $i \in \{0, 1, \ldots, n-1\}$, inequality (5.12) is satisfied with equality at a given $\boldsymbol{x} \in X'$ if and only if *at least* i entries of $\boldsymbol{x}$ realize value r and *at least* $(n-i-1)$ entries realize value 1, due to Theorem 3.1.2. In addition, *every* inequality of (5.16) is satisfied with equality at only the

two points $\mathbb{1}$ and $r\mathbb{1}$. Thus, when $\log_r(q)$ is not an integer, the maximum error is realized at only the $\binom{n}{i^*}+2$ extreme points of X' given by $\mathbb{1}$, $r\mathbb{1}$, and the $\binom{n}{i^*}$ points containing exactly i^* entries of value r and $(n-i^*)$ entries of value 1. However, when $\log_r(q)$ is an integer, the maximum error is realized at $\mathbb{1}$, $r\mathbb{1}$, and every $\boldsymbol{x}\in X'$ such that at least i^* entries of $\boldsymbol{x}$ realize value r and at least $(n-i^*-1)$ entries realize value 1.

The proof is complete. □

Example 8 below illustrates Theorem 5.0.4 and its proof when $\log_r\left(\frac{r^n-1}{n(r-1)}\right)$ is not an integer.

Example 8. *Consider the monomial* $\prod_{j=1}^{n}x_j$ *in* $n=3$ *variables when* $L_1=L_2=L_3=1$, *and* $U_1=U_2=U_3=4$ *in* (1.4) *so that the conditions of Theorem 5.0.4 are satisfied with* $r=4$. *Theorem 5.0.4 gives the error* $\eta(X')$ *of* (5.1) *as* $\eta(X')=\frac{27}{2}$ *with* $L_{(\alpha_0,\boldsymbol{\alpha})}=7(x_1+x_2+x_3)-\frac{67}{2}$, $i^*=2$, *and* $q=\left(\frac{4^3-1}{3(4-1)}\right)=7$. *Since* $log_4(7)$ *is not an integer, the error of* $\frac{27}{2}$ *occurs at only the five points* $(x_1,x_2,x_3)=(1,1,1),(4,4,4),(1,4,4),(4,1,4)$, *and* $(4,4,1)$. *Here, the inequalities* (5.13) *and* (5.14) *from* (5.12) *take the form*

$$4(x_1+x_2+x_3-5)\le x_1x_2x_3 \quad \textit{and} \quad 16(x_1+x_2+x_3-8)\le x_1x_2x_3,$$

and are surrogated using $\frac{3}{4}$ *and* $\frac{1}{4}$, *respectively, to obtain*

$$-27\le x_1x_2x_3-7(x_1+x_2+x_3)+20,$$

which is the left inequality of (5.11). *Also, inequalities* (5.16) *take the form*

$$\begin{aligned} x_1x_2x_3&\le 16x_1+4x_2+x_3-20, & x_1x_2x_3&\le 16x_1+x_2+4x_3-20,\\ x_1x_2x_3&\le 4x_1+16x_2+x_3-20, & x_1x_2x_3&\le 4x_1+x_2+16x_3-20,\\ x_1x_2x_3&\le x_1+16x_2+4x_3-20, & x_1x_2x_3&\le x_1+4x_2+16x_3-20, \end{aligned}$$

and are surrogated using multipliers of $\frac{1}{6}$ *each to obtain*

$$x_1x_2x_3-7(x_1+x_2+x_3)+20\le 0$$

as the right inequality of (5.11).

Example 9 illustrates Theorem 5.0.4 and its proof when $\log_r\left(\frac{r^n-1}{n(r-1)}\right)$ is an integer.

Example 9. *Consider the monomial* $\prod_{j=1}^{n} x_j$ *in* $n = 4$ *variables when* $L_1 = L_2 = L_3 = L_4 = 1$, *and* $U_1 = U_2 = U_3 = U_4 = \left(1+\sqrt{2}\right)$ *in* (1.4) *so that the conditions of Theorem 5.0.4 are satisfied with* $r = \left(1+\sqrt{2}\right)$. *Theorem 5.0.4 gives the error* $\eta(X')$ *of* (5.1) *as* $\eta(X') = \left(3+2\sqrt{2}\right)$ *with* $L_{(\alpha_0,\boldsymbol{\alpha})} = \left(3+2\sqrt{2}\right)(x_1+x_2+x_3+x_4) - \left(14+10\sqrt{2}\right)$, $i^* = 2$, *and* $q = \left(\frac{(1+\sqrt{2})^4-1}{4(1+\sqrt{2}-1)}\right) = (3+2\sqrt{2})$. *Since* $log_{\left(1+\sqrt{2}\right)}\left(3+2\sqrt{2}\right) = 2$ *is an integer, the error of* $\left(3+2\sqrt{2}\right)$ *occurs at the points* $(x_1,x_2,x_3,x_4) = (1,1,1,1)$, $\left(1+\sqrt{2},1+\sqrt{2},1+\sqrt{2},1+\sqrt{2}\right)$, *and every* $\boldsymbol{x} \in X'$ *such that at least 2 entries of* $\boldsymbol{x}$ *realize value* $\left(1+\sqrt{2}\right)$ *and at least 1 entry realizes value 1. The inequality* (5.15) *from* (5.12) *can be expressed as*

$$-6-4\sqrt{2} \leq x_1x_2x_3x_4 - \left(3+2\sqrt{2}\right)(x_1+x_2+x_3+x_4) + 11 + 8\sqrt{2},$$

which is the left inequality of (5.11). *Also, the* $4! = 24$ *inequalities of* (5.16) *are surrogated using multipliers of* $\frac{1}{24}$ *each to obtain*

$$x_1x_2x_3x_4 - \left(3+2\sqrt{2}\right)(x_1+x_2+x_3+x_4) + 11 + 8\sqrt{2} \leq 0$$

as the right inequality of (5.11).

Lemma 5.0.1 allows for a generalization of Theorem 5.0.4 that identifies an optimal solution to (5.1) when, for any given $r > 1$, each j has either $L_j > 0$ and $U_j = rL_j$ or has $U_j < 0$ and $L_j = rU_j$. For such cases, we can set $c_j = L_j$ if $L_j > 0$ and $c_j = U_j$ if $U_j < 0$ to obtain Corollary 5.3.1 below from Theorem 5.0.4.

Corollary 5.3.1. *For any given* $r > 1$, *when each* j *has either* $L_j > 0$ *and* $U_j = rL_j$ *or has* $U_j < 0$ *and* $L_j = rU_j$, *the error* $\eta(X'_C)$ *of* (5.2) *is* $\eta(X'_C) = \left|\prod_{i=1}^{n} c_i\right| \left[\frac{1}{2}\left(1+(r-1)qi^* - r^{i^*}\right)\right]$, *and there exists a unique linear function affording this error, defined with* $\alpha_0 = -\left(\prod_{i=1}^{n} c_i\right)\left[\frac{1}{2}\left(2nq+(r-1)qi^* - r^{i^*} - 1\right)\right]$ *and* $\alpha_j = \left(\prod_{i\neq j} c_i\right) q$ *for all* j, *so that*

$$L_{(\alpha_0,\boldsymbol{\alpha})} = \sum_{j=1}^{n}\left(\prod_{i\neq j} c_i\right) qx_j - \left(\prod_{i=1}^{n} c_i\right)\left[\frac{1}{2}\left(2nq+(r-1)qi^* - r^{i^*} - 1\right)\right].$$

Here, $i^* = \lceil log_r(q) \rceil$ *with* $q = \left(\frac{r^n-1}{n(r-1)}\right)$, *and* $\boldsymbol{c}$ *has* $c_j = L_j$ *if* $L_j > 0$ *and* $c_j = U_j$ *if* $U_j < 0$.

Moreover, when $\log_r(q)$ is not an integer, this maximum error is realized at only the $\binom{n}{i^*}+2$ extreme points of X'_C given by $\boldsymbol{c}$, $r\boldsymbol{c}$, and the $\binom{n}{i^*}$ points which coincide with exactly i^* entries of $r\boldsymbol{c}$ and $(n-i^*)$ entries of $\boldsymbol{c}$. However, when $\log_r(q)$ is an integer, the maximum error is realized at $\boldsymbol{c}$, $r\boldsymbol{c}$, and every $\boldsymbol{x} \in X'_C$ which coincides with at least i^* entries of $r\boldsymbol{c}$ and at least $(n-i^*-1)$ entries of $\boldsymbol{c}$.

Example 10 below illustrates Corollary 5.3.1.

Example 10. *Consider the monomial $\prod_{j=1}^{n} x_j$ in $n = 3$ variables when $L_1 = 1$, $L_2 = -12$, $L_3 = 2$, $U_1 = 4$, $U_2 = -3$, and $U_3 = 8$ in (1.4) so that the conditions of Corollary 5.3.1 are satisfied with $r = 4$. Corollary 5.3.1 gives the error $\eta(X'_C)$ of (5.2) as $\eta(X'_C) = 81$ with $L_{(\alpha_0,\boldsymbol{\alpha})} = -42x_1 + 14x_2 - 21x_3 + 201$, $i^* = 2$, $q = \left(\frac{4^3-1}{3(4-1)}\right) = 7$, and $\boldsymbol{c} = (1,-3,2)$. In addition, since $\log_4(7)$ is not an integer, Corollary 5.3.1 states that this error occurs at only the five points $(x_1,x_2,x_3) = (1,3,-2),(4,-12,8),(1,-12,8),(4,-3,8)$, and $(4,-12,-2)$.*

5.4 Proof of Theorem 5.0.5, where $X' = X = [-1,1]^n$

Beyond the cases found within Theorem 5.0.2 and Corollary 5.2.1, the error $\eta(X')$ of (5.1) can be computed for those instances of $\prod_{j=1}^{n} x_j$ for which (1.4) has $L_j = -1$ and $U_j = 1$ for all j. These instances are addressed in Theorem 5.0.5 below.

Proof of Theorem 5.0.5. Following the proofs of Theorems 5.0.2 and 5.0.3, the argument consists of three parts: the first part establishes 1 as a lower bound on $\eta(X')$, the second part shows that the only linear function $L_{(\alpha_0,\boldsymbol{\alpha})}$ that can possibly yield this bound is defined in terms of the stated $(\alpha_0,\boldsymbol{\alpha})$, and the third part shows that this bound is indeed realized for the given $L_{(\alpha_0,\boldsymbol{\alpha})}$ at only the 2^n extreme points of $X' = [-1,1]^n$.

- Let $\boldsymbol{x}^j$ for $j = 1,\ldots,2^n$ denote the 2^n extreme points of X', with x_k^j representing entry k of

extreme point $\boldsymbol{x}^j$. Regardless of the values of $(\alpha_0, \boldsymbol{\alpha})$, the 2^n extreme points of X' give us that

$$
\begin{aligned}
&(2^n)\left\{\max_{\boldsymbol{x}\in X'}\left|\prod_{j=1}^{n} x_j - (\boldsymbol{\alpha}^T\boldsymbol{x}+\alpha_0)\right|\right\} \geq \sum_{j=1}^{2^n}\left|P^j - \left(\sum_{k=1}^{n}\alpha_k x_k^j + \alpha_0\right)\right| \\
&= \sum_{j=1}^{2^n}\left|P^j\left[P^j - \left(\sum_{k=1}^{n}\alpha_k x_k^j + \alpha_0\right)\right]\right| \geq \sum_{j=1}^{2^n} P^j\left[P^j - \left(\sum_{k=1}^{n}\alpha_k x_k^j + \alpha_0\right)\right] \\
&= \sum_{j=1}^{2^n}\left[1 - P^j\left(\sum_{k=1}^{n}\alpha_k x_k^j\right) - P^j\alpha_0\right] = 2^n - \sum_{j=1}^{2^n}\left[P^j\left(\sum_{k=1}^{n}\alpha_k x_k^j\right)\right] - 0 \\
&= 2^n - \sum_{k=1}^{n}\alpha_k\left[\sum_{j=1}^{2^n} P^j x_k^j\right] = 2^n - \sum_{k=1}^{n}\alpha_k\left[\sum_{j=1}^{2^n}\left(\prod_{l\neq k} x_l^j\right)\right] \\
&= 2^n. \qquad (5.17)
\end{aligned}
$$

Then (5.1) has $2^n(\eta(X')) \geq 2^n$ so that $\eta(X') \geq 1$. Here, for each $j = 1, \ldots, 2^n$, the scalar P^j is either -1 or 1, defined as $P^j = \left(\prod_{l=1}^{n} x_l^j\right)$.

- Suppose there exists an $(\alpha_0, \boldsymbol{\alpha})$ so that $\eta(X')$ from (5.1) has $\eta(X') = 1$. Then both inequalities of (5.17) must be satisfied at equality for this $(\alpha_0, \boldsymbol{\alpha})$. Since each absolute value expression in the first inequality would be bounded above by $\eta(X')$, this inequality would be satisfied at equality if and only if each of these 2^n expressions is equal to 1. But then the second inequality would also be satisfied at equality if and only if each of the following 2^n equations holds:

$$
\begin{aligned}
1 - \left(\sum_{k=1}^{n}\alpha_k x_k^j + \alpha_0\right) = 1 \;\forall\, j = 1, \ldots, 2^n \text{ with } P^j = 1 \\
\text{and } 1 + \left(\sum_{k=1}^{n}\alpha_k x_k^j + \alpha_0\right) = 1 \;\forall\, j = 1, \ldots, 2^n \text{ with } P^j = -1,
\end{aligned}
$$

so that

$$\sum_{k=1}^{n}\alpha_k x_k^j + \alpha_0 = 0 \;\forall\, j = 1, \ldots, 2^n. \qquad (5.18)$$

Now insert that extreme point, say $\boldsymbol{x}^p$, having $x_k^p = 1$ for all k into (5.18) to obtain

$$\sum_{k=1}^{n}\alpha_k + \alpha_0 = 0. \qquad (5.19)$$

Next, given any $s \in \{1, \ldots, n\}$, insert that extreme point, say $\boldsymbol{x}^q$, having $x_s^q = -1$ and $x_j^q = 1$

for all $j \neq s$ into (5.18) to obtain $-\alpha_s + \sum_{k \neq s} \alpha_k + \alpha_0 = 0$. Subtracting this equation from (5.19), we obtain $2\alpha_s = 0$. Since then all $\alpha_s = 0$, equation (5.19) gives us that $\alpha_0 = 0$.

- Let $(\alpha_0, \boldsymbol{\alpha})$ be as stated in the Theorem and derived above. Then the inner maximization problem of (5.1) is

$$\max_{\boldsymbol{x} \in X'} \left\{ \left| \prod_{j=1}^{n} x_j \right| \right\}.$$

The optimal objective value is clearly 1, and this value is realized at only the 2^n extreme points of X'.

□

Lemma 5.0.1 permits a generalization of Theorem 5.0.5 that identifies an optimal solution to (5.1) when $L_j = -U_j$ for all j. This generalization follows by letting $c_j = U_j$ for all j. The formal statement is given below.

Corollary 5.4.1. *When $L_j = -U_j$ for all j, the error $\eta(X'_C)$ of (5.2) is $\eta(X'_C) = \left(\prod_{j=1}^{n} U_j \right)$, and there exists a unique linear function affording this error, defined with $\alpha_0 = 0$ and $\alpha_j = 0$ for all j, so that*

$$L_{(\alpha_0, \boldsymbol{\alpha})} = \sum_{j=1}^{n} 0x_j + 0.$$

Moreover, this maximum error is realized at only the 2^n extreme points of X'_C.

Example 11. *Consider the monomial $\prod_{j=1}^{n} x_j$ in $n = 3$ variables when $L_1 = -U_1 = -2$, $L_2 = -U_2 = -3$, and $L_4 = -U_4 = -4$ in (1.4). Corollary 5.4.1 gives the error $\eta(X'_C)$ of (5.2) as $\eta(X'_C) = 24$, which is associated with the linear function $L_{(\alpha_0, \boldsymbol{\alpha})} = 0x_1 + 0x_2 + 0x_3 + 0$. In addition, Corollary 5.4.1 states that this error occurs at only the eight points $(x_1, x_2, x_3) = (-2, -3, -4), (-2, -3, 4), (-2, 3, -4), (-2, 3, 4), (2, -3, -4), (2, -3, 4), (2, 3, -4)$, and $(2, 3, 4)$.*

5.5 Discussions on Error Comparison

Finally, we perform the error comparison between the worst-case errors obtained for the two approximation approaches established in Chapter 4 and 5:

1. When $n = 2$, $\mu(X') = \eta(X') = \frac{(U_1 - L_1)(U_2 - L_2)}{4}$. Recall the discussion before Example 5.

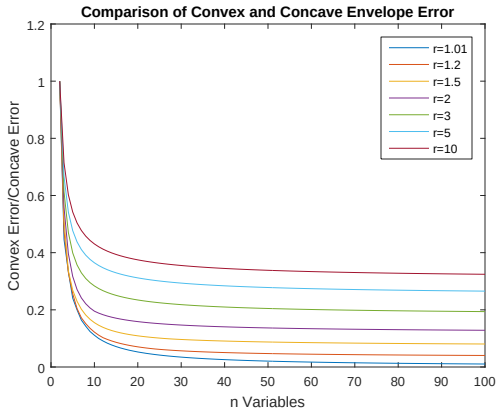

Figure 5.1: $\mathcal{D}_{r,n}/\mathcal{E}_{r,n}$

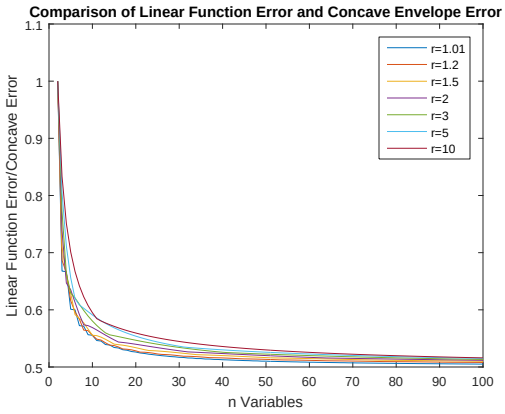

Figure 5.2: $\mathcal{F}_{r,n}/\mathcal{E}_{r,n}$

2. When $X' = [0,1]^n$, and $n \geq 3$, $\left(1-\frac{1}{n}\right) n^{\frac{1}{1-n}} = \mu(X') > \eta(X') = \frac{n-1}{2n}$, provided by the following:

   Proposition 5.5.1. $\left(1-\frac{1}{n}\right) n^{\frac{1}{1-n}} > \frac{n-1}{2n}$ *for* $n \geq 3$.

   Proof. It is equivalent to show that $n^{\frac{1}{1-n}} > \frac{1}{2}$, which is equivalent to $2 > n^{\frac{1}{n-1}}$, then to $2^{n-1} > n$. □

3. When $X' = [1,r]^n$, and $n \geq 3$, we provide the following asymptotic results: as n gets large,

   - $\mathcal{E}_{r,n}/r^n \to 1$, which is easy to check once we establish $\left(\frac{r^n-1}{n(r-1)}\right)^{\frac{1}{n-1}} \to r$;
   - $\lim_{n\to\infty} \mathcal{D}_{r,n}/r^n < 1/e$, given by Remark 7;
   - $\mathcal{F}_{r,n}/r^n \to 1/2$, which is easy to check once we establish $i^*/n \to 1$ and $r^{i^*}/r^n \to 0$: the former utilizes $\left(\frac{r^n-1}{n(r-1)}\right)^{\frac{1}{n}} \to r$, and the latter utilizes $r^{i^*} \approx \frac{r^n-1}{n(r-1)}$.

   The strict inequality that $\max\{\mathcal{D}_{r,n}, \mathcal{E}_{r,n}\} = \mu(X') > \eta(X') = \mathcal{F}_{r,n}$ is conjectured, proved asymptotically, and checked computationally. Figure 5.1 and Figure 5.2 illustrate the comparisons between the Convex Envelope Error ($\mu^{\text{vex}} = \mathcal{D}_{r,n}$), the Concave Envelope Error ($\mu^{\text{cav}} = \mathcal{E}_{r,n}$) and the Linear Function Error ($\eta = \mathcal{F}_{r,n}$), for various r values chosen, when n goes from 2 to 100.

   Notice that when $n = 2$, the Convex Envelope Error coincides with the Concave Envelope Error, however the gap grows rapidly as n gets larger. Recall that the maximum Convex Envelope Error and maximum Concave Envelope Error both occur on the segment between

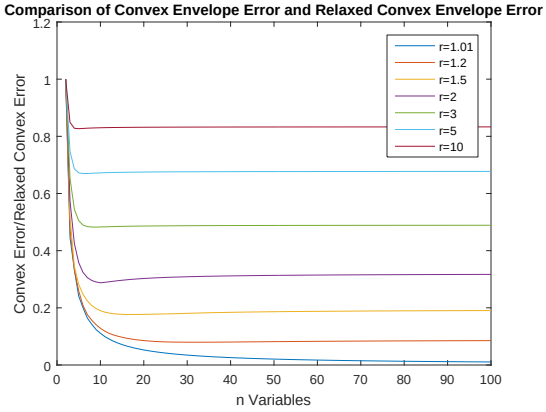

Figure 5.3: Relaxation Quality in the Convex Envelope

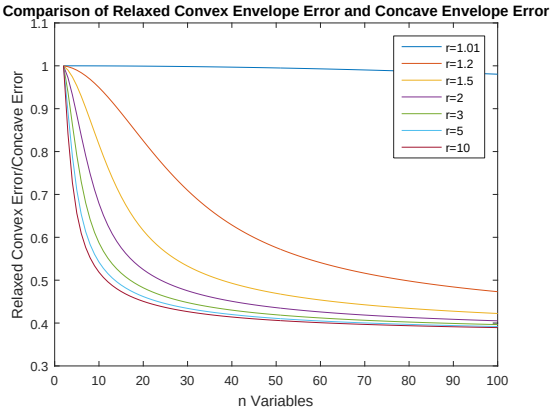

Figure 5.4: Relaxed Convex Envelope Strength

$\mathbb{1}, r\mathbb{1}$, therefore we can further conjecture that, if we relax the under-estimator from (4.10) to consist of only the 1st and nth facet ($i = 1, n$), it would still beat the Concave Envelope in the worst case. This conjecture is illustrated in Figure 5.3 and Figure 5.4.

4. When $X' = [-1, 1]^n$, and $n \geq 3$, $1 + \left(\frac{n-2}{n}\right)^n = \mu\left(X'\right) > \eta\left(X'\right) = 1$.

For certain formulations of the optimization problem, if the concave envelope inequalities are redundant, then the worst case error produced by the concave envelope will never be realized, one may think about the example where $\prod_{i=1}^{n} x_i$ is minimized only, in the objective function, and only the worst case error produced by the convex envelope comes into play.

1. For the $[0, 1]^n$ box, one can see that the Convex Envelope Error ($\mu^{\text{vex}} =$) $\left(\frac{n-1}{n}\right)^n$ is always upper bounded by the Linear Function Error ($\eta =$) $\frac{n-1}{2n}$, because the comparison essentially leads to $2 \leq \left(\frac{n}{n-1}\right)^{n-1}$, which is true under the same logic of (4.19). Asymptotically, the former tends to e^{-1}, yet the latter tends to $\frac{1}{2}$.

2. For the $[1, r]^n$ box, one can see that the Convex Envelope Error ($\mu^{\text{vex}} = \mathcal{D}_{r,n}$) is always less than the Linear Function Error ($\eta = \mathcal{F}_{r,n}$) asymptotically, that $\lim_{n\to\infty} \mathcal{D}_{r,n}/\mathcal{E}_{r,n} < 1/e$ and $\lim_{n\to\infty} \mathcal{F}_{r,n}/\mathcal{E}_{r,n} = 1/2$, and graphically from Figure 5.1 and Figure 5.2: $\mathcal{D}_{r,n}/\mathcal{E}_{r,n}$ rapidly gets below the .4 level, while $\mathcal{F}_{r,n}/\mathcal{E}_{r,n}$ stays above the .5 level.

3. For the $[-1, 1]^n$ box, one should notice that the Convex Envelope Error (μ^{vex}) is actually the same as the Concave Envelope Error (μ^{cav}), in the sense that the convex hull is symmetric

with respect to reflections. On this box, the Linear Function Error (η) still beats both the Convex and Concave Envelope in the worst case.

Bibliography

[1] Warren Adams, Akshay Gupte, and Yibo Xu. Error bounds for monomial convexification in polynomial optimization. *arXiv preprint arXiv:1704.00424*, 2017.

[2] Faiz A. Al-Khayyal and James E. and Falk. Jointly constrained biconvex programming. *Math. Oper. Res.*, 8(2):273–286, 1983.

[3] Martin Ballerstein and Dennis Michaels. Extended formulations for convex envelopes. *J. Global Optim.*, 60(2):217–238, 2014.

[4] Harold P. Benson. Concave envelopes of monomial functions over rectangles. *Naval Res. Logist.*, 51(4):467–476, 2004.

[5] Natashia Boland, Santanu S. Dey, Thomas Kalinowski, Marco Molinaro, and Fabian Rigterink. Bounding the gap between the McCormick relaxation and the convex hull for bilinear functions. *Math. Program.*, 162(1-2, Ser. A):523–535, 2017.

[6] Endre Boros and Peter L. Hammer. Pseudo-Boolean optimization. *Discrete Appl. Math.*, 123(1-3):155–225, 2002. Workshop on Discrete Optimization, DO'99 (Piscataway, NJ).

[7] Samuel Burer and Adam N. Letchford. On nonconvex quadratic programming with box constraints. *SIAM J. Optim.*, 20(2):1073–1089, 2009.

[8] C. Carathéodory. Über den Variabilitätsbereich der Koeffizienten von Potenzreihen, die gegebene Werte nicht annehmen. *Math. Ann.*, 64(1):95–115, 1907.

[9] Yves Crama. Recognition problems for special classes of polynomials in 0-1 variables. *Math. Program.*, 44(2, (Ser. A)):139–155, 1989.

[10] Yves Crama. Concave extensions for nonlinear 0-1 maximization problems. *Math. Program.*, 61(1, Ser. A):53–60, 1993.

[11] Pasquale L. De Angelis, Panos M. Pardalos, and Gerardo Toraldo. Quadratic programming with box constraints. In *Developments in Global Optimization (Szeged, 1995)*, volume 18 of *Nonconvex Optim. Appl.*, pages 73–93. Kluwer Acad. Publ., Dordrecht, 1997.

[12] Etienne de Klerk and Monique Laurent. Error bounds for some semidefinite programming approaches to polynomial minimization on the hypercube. *SIAM J. Optim.*, 20(6):3104–3120, 2010.

[13] Etienne de Klerk, Monique Laurent, and Zhao Sun. An error analysis for polynomial optimization over the simplex based on the multivariate hypergeometric distribution. *SIAM J. Optim.*, 25(3):1498–1514, 2015.

[14] Etienne de Klerk, Monique Laurent, and Zhao Sun. Convergence analysis for Lasserre's measure-based hierarchy of upper bounds for polynomial optimization. *Math. Program.*, 162(1-2, Ser. A):363–392, 2017.

[15] Alberto Del Pia and Aida Khajavirad. A polyhedral study of binary polynomial programs. *Math. Oper. Res.*, 42(2):389–410, 2017.

[16] Fred Glover. Improved linear integer programming formulations of nonlinear integer problems. *Management Sci.*, 22(4):455–460, 1975/76.

[17] Fred Glover and Eugene Woolsey. Further reduction of zero-one polynomial programming problems to zero-one linear programming problems. *Operations Res.*, 21:156–161, 1973. Mathematical programming and its applications.

[18] Fred Glover and Eugene Woolsey. Converting the 0-1 polynomial programming problem to a 0-1 linear program. *Operations Research*, 22(1):180–182, 1974.

[19] James Luedtke, Mahdi Namazifar, and Jeff Linderoth. Some results on the strength of relaxations of multilinear functions. *Math. Program.*, 136(2, Ser. B):325–351, 2012.

[20] Garth P. McCormick. Computability of global solutions to factorable nonconvex programs. I. Convex underestimating problems. *Math. Program.*, 10(2):147–175, 1976.

[21] Clifford A. Meyer and Christodoulos A. Floudas. Trilinear monomials with mixed sign domains: facets of the convex and concave envelopes. *J. Global Optim.*, 29(2):125–155, 2004.

[22] Clifford A. Meyer and Christodoulos A. Floudas. Convex envelopes for edge-concave functions. *Math. Program.*, 103(2, Ser. B):207–224, 2005.

[23] Trang T Nguyen, Jean-Philippe P Richard, and Mohit Tawarmalani. Deriving convex hulls through lifting and projection. *Math. Program.*, pages 1–39, 2017.

[24] Jong-Shi Pang. Error bounds in mathematical programming. *Math. Program.*, 79(1-3, Ser. B):299–332, 1997. Lectures on mathematical programming (ismp97) (Lausanne, 1997).

[25] Anatoliy D. Rikun. A convex envelope formula for multilinear functions. *J. Global Optim.*, 10(4):425–437, 1997.

[26] Hong Seo Ryoo and Nikolaos V. Sahinidis. Analysis of bounds for multilinear functions. *J. Global Optim.*, 19(4):403–424, 2001.

[27] Hanif D. Sherali. Convex envelopes of multilinear functions over a unit hypercube and over special discrete sets. *Acta Math. Vietnam.*, 22(1):245–270, 1997.

[28] Hanif D. Sherali and Warren P. Adams. A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. *SIAM J. Discrete Math.*, 3(3):411–430, 1990.

[29] Hanif D. Sherali and Warren P. Adams. A hierarchy of relaxations and convex hull characterizations for mixed-integer zero-one programming problems. *Discrete Appl. Math.*, 52(1):83–106, 1994.

[30] Hanif D. Sherali and Warren P. Adams. *A reformulation-linearization technique for solving discrete and continuous nonconvex problems*, volume 31 of *Nonconvex Optimization and its Applications*. Kluwer Academic Publishers, Dordrecht, 1999.

[31] Emily Speakman and Jon Lee. Quantifying double McCormick. *Math. Oper. Res.*, 42(4):1230–1253, 2017.

[32] Fabio Tardella. Existence and sum decomposition of vertex polyhedral convex envelopes. *Optim. Lett.*, 2(3):363–375, 2008.

[33] Mohit Tawarmalani, Jean-Philippe P. Richard, and Chuanhui Xiong. Explicit convex and concave envelopes through polyhedral subdivisions. *Math. Program.*, 138(1-2, Ser. A):531–577, 2013.

[34] Yibo Xu. Convex hull derivation for a symmetric multilinear polynomial and a symmetric polytope. *working paper*, 2018.

[35] Yibo Xu, Warren Adams, and Akshay Gupte. Deriving convex hull forms of special symmetric multilinear polynomials. *working paper*, 2018.