Noname

Some Results on the Strength of Relaxations of Multilinear Functions

James Luedtke · Mahdi Namazifar · Jeff Linderoth

the date of receipt and acceptance should be inserted later

Abstract We study approaches for obtaining convex relaxations of global optimization problems containing multilinear functions. Specifically, we compare the concave and convex envelopes of these functions with the relaxations that are obtained with a standard relaxation approach, due to McCormick. The standard approach reformulates the problem to contain only bilinear terms and then relaxes each term independently. We show that for a multilinear function having a single product term, this approach yields the convex and concave envelopes if the bounds on all variables are symmetric around zero. We then review and extend some results on conditions when the concave envelope of a multilinear function can be written as a sum of concave envelopes of its individual terms. Finally, for bilinear functions we prove that the difference between the concave upper bounding and convex lower bounding functions obtained from the McCormick relaxation approach is always within a constant of the difference between the concave and convex envelopes. These results, along with numerical examples we provide, give insight into how to construct strong relaxations of multilinear functions.

Keywords Global optimization · Bilinear function · Multilinear function

1 Introduction

The construction of convex lower bounding and concave upper bounding functions for nonconvex functions plays a critical role in algorithms for globally solving nonconvex optimiza-

This research was supported in part by the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy under Grant DE-FG02-08ER25861.

J. Luedtke · M. Namazifar · J. Linderoth
Department of Industrial and Systems Engineering, University of Wisconsin-Madison, USA
E-mail: jrluedt1@wisc.edu

M. Namazifar
E-mail: namazifar@wisc.edu

J. Linderoth
E-mail: linderoth@wisc.edu

tion problems. In this work, we focus on multilinear functions $\phi : [\ell, u] \to \mathbb{R}$, where

$$\phi(x) = \sum_{t \in T} a_t \prod_{j \in J_t} x_j, \tag{1}$$

and $[\ell, u] = \{x \in \mathbb{R}^n \mid \ell \leq x \leq u\}$. Specifically, we are interested in comparing the strength of relaxations of the graph of such a function, given by the set

$$X \stackrel{\text{def}}{=} \{(x, z) \in [\ell, u] \times \mathbb{R} \mid z = \phi(x)\}.$$

An important special case is when ϕ is a *bilinear* function, i.e., $|J_t| \leq 2$ for all $t \in T$.

When $\phi(x)$ consists of a single bilinear term, McCormick [14] proposed to relax the set $B = \{(x_1, x_2, z) \in [\ell_1, u_1] \times [\ell_2, u_2] \times \mathbb{R} \mid z = x_1 x_2\}$ with the following inequalities, which we refer to as the McCormick inequalities:

$$z \geq u_2 x_1 + u_1 x_2 - u_1 u_2, \quad z \geq \ell_2 x_1 + \ell_1 x_2 - \ell_1 \ell_2, \tag{2a}$$

$$z \leq u_2 x_1 + \ell_1 x_2 - \ell_1 u_2, \quad z \leq \ell_2 x_1 + u_1 x_2 - u_1 \ell_2. \tag{2b}$$

Al-Khayyal and Falk [1] showed that the convex hull of B is described by (2). For more general factorable nonconvex functions, including multilinear functions of the form (1), McCormick proposed a recursive procedure in which additional variables and constraints are added to obtain a formulation of the problem having only bilinear equations which are subsequently relaxed using (2). The resulting relaxation, which we refer to as the *McCormick relaxation*, has formed a basis for the relaxations used in many global optimization solution approaches, such as implemented in BARON [21,24], Couenne [3], and [23].

The strongest possible relaxation of X, its convex hull $\mathsf{conv}(X)$, has been shown to be a polyhedron with the following characterization [6–8,19,22]:

$$\mathsf{conv}(X) = \operatorname*{Proj}_{x,z} \left\{ (x, z, \lambda) \in [\ell, u] \times \mathbb{R} \times \Delta_{2^n} \mid x = \sum_{j=1}^{2^n} \lambda_j x^j, z = \sum_{j=1}^{2^n} \lambda_j \phi(x^j) \right\}, \tag{3}$$

where $x^1, x^2, \ldots x^{2^n}$ are the vertices of $[\ell, u]$, and Δ_{2^n} is the 2^n-dimensional simplex. In general, the McCormick relaxation may strictly contain the convex hull, leading to weaker relaxation bounds. On the other hand, direct use of the convex hull characterization (3) to create a convex relaxation of X is limited by the exponential growth in the number of variables. Thus, a natural idea is to seek relaxations of X that may be tighter than what is obtained with the standard McCormick approach, but which are not as prohibitively large as the full convex hull approach. A simple idea along these lines is to use the formulation (3) over *subsets* of the variables chosen small enough to keep the size of the relaxation tractable. This idea has already been explored with promising results by Bao, Sahinidis, and Tawarmalani [2], where procedures to find valid inequalities based on the dual formulation of (3) are investigated. We also refer the reader to the Ph.D. thesis of the second author [17] for a more detailed exposition of some of the results presented in this paper.

Since using (3) in any form is likely to increase the computational burden in solving the relaxation, it is important to understand when this extra work is most likely to yield significant benefits in relaxation quality. To this end, we explore conditions under which the convex hull formulation yields nothing more than McCormick relaxation approach, and, for the case of bilinear functions, we provide bounds on how much worse the McCormick relaxation can be. To our knowledge, this is the first result of this type in the global optimization literature.

We begin in §2 with the case in which ϕ consists of single product term ($|T| = 1$). We first review a result of Ryoo and Sahinidis [20] that shows the McCormick relaxation is equivalent to the convex hull when the bounds on the variables are all $[0, 1]$. We then provide the new result that this also holds when the bounds are symmetric about zero, i.e., $x_i \in [-u_i, u_i]$. Finally, we provide examples that when these conditions do not hold, the difference between the convex hull and McCormick relaxations can be arbitrarily large.

In §3, we consider the case when ϕ can have multiple terms. We begin in §3.1 by reviewing an existing result of Meyer and Floudas [15] which states that the *concave* envelope of ϕ over $x \in [0, 1]^n$ can be obtained as the sum of concave envelopes of the individual terms of ϕ when the coefficients on each term are positive. We show that this result extends to $x \in [\ell, u]$ provided $\ell \in \mathbb{R}^n_+$, and to general $[\ell, u]$ if ϕ is bilinear. Note that these results do not say anything about how the *convex lower bounding function* obtained from the McCormick relaxation compares to the convex envelope.

In §3.2 we focus on bilinear functions of the form $b(x) = \sum_{\{i,j\}\in E} a_{ij} x_i x_j$, where E is a set of unordered pairs of distinct elements of $N = \{1, \ldots, n\}$, and obtain results that provide insight into the strength of the McCormick upper *and* lower bounding functions, relative to the concave and convex envelopes. To motivate these results, consider an experiment comparing the McCormick relaxation to the convex hull for the following two bilinear functions defined on $x \in H = [0, 1]^7$:

$$\begin{aligned} b_1(x) =& x_1x_2 + x_1x_3 + x_1x_4 + x_1x_5 + x_1x_6 + x_2x_3 + x_2x_4 + x_2x_5 + x_3x_4 \\ &+ x_3x_5 + x_4x_5 + x_4x_6 + x_5x_7 + x_6x_7, \\ b_2(x) =& x_1x_2 - x_1x_3 + x_1x_4 + x_1x_5 + x_1x_6 + x_2x_3 - x_2x_4 - x_2x_5 + x_3x_4 \\ &+ x_3x_5 - x_4x_5 + x_4x_6 - x_5x_7 + x_6x_7. \end{aligned}$$

For each of these functions, we randomly generated 5000 points uniformly in H and, for each point x^k, calculated the difference between the concave and convex envelopes of b at x^k, denoted $\mathsf{chgap}_H[b](x^k)$, and also calculated the difference between the upper and lower bounds of $b(x^k)$ defined by the McCormick relaxation, denoted $\mathsf{mcgap}_H[b](x^k)$. (These terms are formally defined in §3.2.) We then construct a scatter plot, shown in Figure 1, of the points $(\mathsf{mcgap}_H[b](x^k), \mathsf{chgap}_H[b](x^k)), k = 1, \ldots, 5000$. Because the McCormick relaxation is weaker than the convex hull, it always holds that $\mathsf{mcgap}_H[b](x) \geq \mathsf{chgap}_H[b](x)$ and so all of these points lie below the line of slope one passing through the origin. The distance of the points from this line provides a graphical illustration of the quality of the McCormick relaxation at each point. A surprising feature of these plots is that all points lie above a line having smaller slope, suggesting that there exists a constant $C_b \geq 1$, depending on the bilinear function b, such that $\mathsf{mcgap}_H[b](x)/\,\mathsf{chgap}_H[b](x) \leq C_b$ holds for all $x \in H$. In §3.2, we prove that this is indeed the case, and furthermore we provide bounds on the approximation constant C_b. If $a_{ij} > 0$ for all $\{i, j\} \in E$, this constant is always less than 2, and decreases with the coloring number of the graph $G = (N, E)$. This yields, as a special case, the result of [5, 9] that the McCormick relaxation is equivalent to the convex hull when G is bipartite. When the coefficients are not all positive, our bound on C_b is $O(n)$. We also show that for any bilinear function, as terms are removed the difference between the convex hull and McCormick relaxation gaps decreases, suggesting that the improvement in relaxation quality by using the convex hull formulation will be more significant when the graph G is denser.

In §4 we present numerical examples that show our results are tight, and also provide insights into the gap between these relaxations for cases where our results do not apply. We make some concluding remarks in §5.

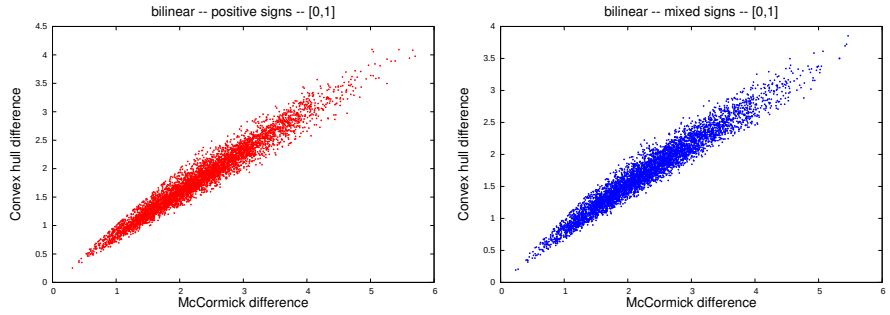

Fig. 1 Scatter plots of McCormick gap vs. convex hull gap for random points in $[0,1]^7$ for a bilinear function having positive coefficients (left) and mixed-sign coefficients (right).

Notation: Given a function $f : D \to \mathbb{R}$, the concave envelope of f over D, written $\mathsf{cav}_D[f]$, is the minimum concave upper bounding function of f on D. That is, $\mathsf{cav}_D[f](x) \geq f(x)$ for all $x \in D$, and if $g : D \to \mathbb{R}$ is any other concave function with $g \geq f$ on D, then $g \geq \mathsf{cav}_D[f]$. Similarly, the convex envelope of f over D, written $\mathsf{vex}_D[f]$, is the maximum convex lower bounding function of f on D. We let $\mathbf{0}$ and $\mathbf{1}$ denote vectors of all zeros and all ones, and e_i be a vector of all zeros except the ith component which has value 1. The lengths of $\mathbf{0}$,$\mathbf{1}$, and e_i will be clear from context (but are usually all n). We let $H = [\mathbf{0}, \mathbf{1}]$ be the unit hypercube. For $u \in \mathbb{R}^n$, we define $\mathrm{Diag}(u)$ to be the $n \times n$ diagonal matrix with $\mathrm{Diag}(u)_{ii} = u_i$.

2 Recursive McCormick relaxation of a single multilinear term

In this section, we consider a multilinear function consisting of a single term, $f(x) = \prod_{j=1}^{n} x_j$. Specifically, we compare relaxations of set

$$X_{[\ell,u]} = \{(x, y_1) \in [\ell, u] \times \mathbb{R} \mid y_1 = f(x)\}.$$

We consider cases in which a *recursive McCormick relaxation*, constructed by recursively applying the McCormick relaxation to products of pairs of variables, is as strong as $\mathsf{conv}(X)$.

2.1 Preliminaries

We first formally define the recursive McCormick relaxation of the set $X_{[\ell,u]}$. This relaxation is referred to as a *recursive Arithmetic Interval* in [20]. First, for fixed intervals $[\ell_1, u_1]$ and $[\ell_2, u_2]$, define the set $\mathsf{MC}_{[\ell_1,u_1]\times[\ell_2,u_2]}$ to be the set of $(y, x_1, x_2) \in \mathbb{R} \times [\ell_1, u_1] \times [\ell_2, u_2]$ that satisfy the McCormick inequalities (2):

$$\begin{aligned} y &\geq u_2 x_1 + u_1 x_2 - u_1 u_2, \quad & y &\geq \ell_2 x_1 + \ell_1 x_2 - \ell_1 \ell_2, \\ y &\leq u_2 x_1 + \ell_1 x_2 - \ell_1 u_2, \quad & y &\leq \ell_2 x_1 + u_1 x_2 - u_1 \ell_2. \end{aligned}$$

Now suppose $\ell, u \in \mathbb{R}^n$ with $\ell \leq u$. A relaxation of this nonconvex set $X_{[\ell,u]}$ can be constructed in a higher-dimensional space by introducing variables $y_2, \ldots, y_n$ that satisfy

$y_i = x_i y_{i+1}$ for $i = 1, \ldots, n-1$ and $y_n = x_n$ and then relaxing these bilinear constraints with the McCormick inequalities. This leads to a "recursive" McCormick relaxation of $X_{[\ell,u]}$ which is the polytope defined by:

$$\mathsf{RMC}\big(X_{[\ell,u]}\big) = \Big\{(x,y) \in [\ell,u] \times \mathbb{R}^n \mid y_n = x_n, \\ (y_i, x_i, y_{i+1}) \in \mathsf{MC}_{[\ell_i,u_i]\times[\tilde{\ell}_{i+1},\tilde{u}_{i+1}]}, i = 1,\ldots,n-1\Big\}$$

where $\tilde{\ell}_n \stackrel{\text{def}}{=} \ell_n$ and $\tilde{u}_n \stackrel{\text{def}}{=} u_n$ and $\tilde{\ell}_i = \min\{\tilde{u}_{i+1}u_i, \tilde{u}_{i+1}\ell_i, \tilde{\ell}_{i+1}u_i, \tilde{\ell}_{i+1}\ell_i\}$ and $\tilde{u}_i = \max\{\tilde{u}_{i+1}u_i, \tilde{u}_{i+1}\ell_i, \tilde{\ell}_{i+1}u_i, \tilde{\ell}_{i+1}\ell_i\}$ are implied lower and upper bounds on y_i for $i = n-1, \ldots, 1$. The variable y_n could be eliminated from the description of $\mathsf{RMC}\big(X_{[\ell,u]}\big)$, but we include it for notational convenience.

Ryoo and Sahinidis [20] proved the following result.

Theorem 1 ([20]) *Let $f(x) = \prod_{i=1}^n x_i$. The recursive McCormick relaxation describes the convex hull of f over the unit hypercube, i.e., $\mathsf{Proj}_{(x,y_1)}\left(\mathsf{RMC}(X_H)\right) = \mathsf{conv}(X_H)$.*

As observed in [20], when $\ell = \mathbf{0}$, the assumption that $u = \mathbf{1}$ is without loss of generality; i.e., we can show the same result holds for $f(x)$ over $[\mathbf{0}, u]$.

Corollary 1 $\mathsf{Proj}_{(x,y_1)}\left(\mathsf{RMC}(X_{[\mathbf{0},u]})\right) = \mathsf{conv}(X_{[\mathbf{0},u]})$.

Proof We only need to prove $\mathsf{Proj}_{(x,y_1)}\left(\mathsf{RMC}(X_{[\mathbf{0},u]})\right) \subseteq \mathsf{conv}(X_{[\mathbf{0},u]})$. Let $(x', y_1') \in \mathsf{Proj}_{(x,y_1)}\left(\mathsf{RMC}(X_{[\mathbf{0},u]})\right)$, $\overline{y}_1 = \left(\prod_{i=1}^n u_i\right)^{-1} y_1'$, and $D_u = \mathrm{Diag}(u)$. We claim that $\left(D_u^{-1}x', \overline{y}_1\right) \in \mathsf{Proj}_{(x,y_1)}\left(\mathsf{RMC}(X_H)\right)$. Clearly, $D_u^{-1}x' \in H$. Let $y_2', \ldots, y_n'$ be such that $(x', y') \in \mathsf{RMC}(X_{[\mathbf{0},u]})$ and let $\overline{y}_i = y_i'\left(\prod_{j=i}^n u_j\right)^{-1}$, $i = 1, \ldots, n$. Then, it is easy to check that $(D_u^{-1}x', \overline{y}) \in \mathsf{RMC}(X_H)$. Then, because $(D_u^{-1}x', \overline{y}_1) \in \mathsf{Proj}_{(x,y_1)}\left(\mathsf{RMC}(X_H)\right)$ Theorem 1 implies there exists $\lambda \in \Delta_{2^n}$ such that $\sum_k \lambda_k (x^k, y^k) = (D_u^{-1}x', \overline{y}_1)$ where $x^k, k = 1, \ldots, 2^n$ are the vertices of X_H and $y^k = f(x^k)$. This implies $x' = \sum_k \lambda_k D_u x^k$, and $y' = \sum_k \lambda_k f(x^k) \prod_{i=1}^n u_i$. Since $D_u x^k \in [\mathbf{0}, u]$ and $f(D_u x^k) = f(x^k)\prod_{i=1}^n u_i$ for all k this implies (x', y') can be written as a convex combination of points in $X_{[\mathbf{0},u]}$. □

2.2 Symmetric bounds

We now show another, somewhat surprising, case where the recursive McCormick relaxation defines the convex hull of a single multilinear term. Specifically, we show that the two relaxations are the same if the bounds on x are symmetric about zero, i.e., $x \in [-u, u]$ for some $u \in \mathbb{R}^n_+$. We begin with the case in which $x \in [-\mathbf{1}, \mathbf{1}]$. First observe that in this case, the implied bounds on y_i for $i = 1, \ldots, n$ are $[\tilde{l}_i, \tilde{u}_i] = [-1, 1]$. Consequently, the conditions $(y_i, x_i, y_{i+1}) \in \mathsf{MC}_{[-1,1]^2}$ in the definition of $\mathsf{RMC}(X_{[-\mathbf{1},\mathbf{1}]})$ have the form

$$y_i \geq -x_i - y_{i+1} - 1, \quad y_i \geq x_i + y_{i+1} - 1, \\ y_i \leq x_i - y_{i+1} + 1, \quad y_i \leq -x_i + y_{i+1} + 1,$$

for $i = 1, \ldots, n-1$.

The result is based on the following characterization of the extreme points of $\mathsf{RMC}(X_{[-\mathbf{1},\mathbf{1}]})$.

Theorem 2 *If (x, y) is an extreme point of $\mathsf{RMC}(X_{[-\mathbf{1},\mathbf{1}]})$, then $(x, y) \in \{-1, 1\}^{2n}$.*

Proof For any $(c,d) \in \mathbb{R}^{2n}$, we show that the linear program

$$\max_{(x,y)\in \mathsf{RMC}(X_{[-\mathbf{1},\mathbf{1}]})} cx + dy \tag{4}$$

has an optimal solution $(x^*, y^*) \in \{-1,1\}^{2n}$, which establishes the claim.

For $z \in [-1,1]$ and $t \in N = \{1,\dots,n\}$, define

$$\begin{aligned} f_t(z) = \max_{\substack{x_t,\dots,x_n \\ y_t,\dots,y_n}} & \sum_{i=t}^{n} c_i x_i + \sum_{i=t}^{n} d_i y_i \\ & \text{s.t. } (y_i, x_i, y_{i+1}) \in \mathsf{MC}_{[-1,1]^2}, \quad i = t,\dots,n-1, \\ & \quad y_t = z,\ x_n \in [-1,1]. \end{aligned}$$

Then $f_t(z)$ satisfies the following recursive relationship for $t = 1,\dots,n-1$:

$$f_t(z) \;=\; d_t z \;+\; \max_{(x_t,y_{t+1})} \big\{ c_t x_t \;+\; f_{t+1}(y_{t+1}) \;\big|\; (z, x_t, y_{t+1}) \;\in\; \mathsf{MC}_{[-1,1]^2}\,. \big\} \tag{5}$$

We show by induction that $f_t(z)$ is convex for all t. First, $f_n(z) = d_n z + |c_n|$. Now assume $f_{t+1}(z)$ is convex. It follows that the maximum in the expression for $f_t(z)$ given in (5) is attained at an extreme point of the the polyhedron $Q(z)$, given by the set of $(x_t, y_{t+1}) \in [-1,1]^2$ that satisfy

$$\begin{aligned} -x_t - y_{t+1} \le 1 + z, \quad x_t + y_{t+1} \le 1 + z, \\ -x_t + y_{t+1} \le 1 - z, \quad x_t - y_{t+1} \le 1 - z. \end{aligned}$$

It is easy to check that for any $z \in [-1,1]$ the extreme points of $Q(z)$ are $\{(-1,-z),(1,z),(-z,-1),(z,1)\}$. Therefore,

$$f_t(z) = d_t z + \max\big\{c_t + f_{t+1}(z), -c_t + f_{t+1}(-z), -c_t z + f_{t+1}(-1), c_t z + f_{t+1}(1)\big\}.$$

Each of the functions taken in the max is a convex function of z, showing that $f_t(z)$ is convex.

Finally, observe that (4) is equivalent to $\max_{y_1\in[-1,1]} f_1(y_1)$. As $f_1(y_1)$ is convex, there exists a solution to this with $y_1^* \in \{-1,1\}$. Proceeding inductively, assume there is an optimal solution to (5) with $(x_i^*, y_{i+1}^*) \in \{-1,1\}^2$ for $i = 1,\dots,t-1$. For t, recall that for any fixed z, (5) has an extreme point optimal solution among the set $(x_t, y_{t+1}) \in \{(-1,-z),(1,z),(-z,-1),(z,1)\}$. Thus, using $z = y_t^* \in \{-1,1\}$ (from the induction hypothesis) shows there exists $x_t^* \in \{-1,1\}$ and $y_{t+1}^* \in \{-1,1\}$ optimal for (5). □

Theorem 3 *Let $f(x) = \prod_{i=1}^n x_i$. The recursive McCormick relaxation describes the convex hull of f over $[-\mathbf{1},\mathbf{1}]$, i.e.,* $\mathsf{Proj}_{(x,y_1)}\big(\mathsf{RMC}(X_{[-\mathbf{1},\mathbf{1}]})\big) = \mathsf{conv}(X_{[-\mathbf{1},\mathbf{1}]})$.

Proof We only need to prove $\mathsf{Proj}_{(x,y_1)}\big(\mathsf{RMC}(X_{[-\mathbf{1},\mathbf{1}]})\big) \subseteq \mathsf{conv}(X_{[-\mathbf{1},\mathbf{1}]})$. By Theorem 2, if (x,y) is an extreme point of $\mathsf{RMC}(X_{[-\mathbf{1},\mathbf{1}]})$ then $(x,y) \in \{-1,1\}^{2n}$. For each t, it is easily checked that if $(y_t, x_t, y_{t+1}) \in \mathsf{MC}_{[-1,1]^2}$, $x_t, y_t \in \{-1,1\}$, then $y_t = x_t y_{t+1}$, and therefore $y_1 = \prod_{i=1}^n x_i$. Thus, $(x,y_1) \in X_{[-\mathbf{1},\mathbf{1}]}$. This is sufficient to prove the result, since this shows that every point in $\mathsf{Proj}\big(\mathsf{RMC}(X_{[-\mathbf{1},\mathbf{1}]})\big)$ can be written as a convex combination of points in $X_{[-\mathbf{1},\mathbf{1}]}$. □

Using arguments identical to those in the proof of Corollary 1 yields the following generalization.

Corollary 2 *Let $u \in \mathbb{R}^n_+$. Then* $\mathsf{Proj}_{(x,y_1)}\big(\mathsf{RMC}(X_{[-u,u]})\big) = \mathsf{conv}(X_{[-u,u]})$.

2.3 Worst-case examples

We have seen that when either $\ell = \mathbf{0}$ or $\ell = -u$, the recursive McCormick relaxation of $f(x) = \prod_{i=1}^{n} x_i$ is as good as the convex hull relaxation. We now show that when both of these conditions are violated, the recursive McCormick relaxation can be arbitrarily worse than the convex hull. We measure the relative quality of these relaxations by comparing the distance between the minimum and maximum allowable values for y at a point x. Specifically, for a given $D = [\ell, u]$, we define

$$\mathsf{chgap}_D[f](x) = \underbrace{\max\big\{y \,|\, (x,y) \in \mathsf{conv}(X_D)\big\}}_{=\mathsf{cav}_D[f](x)} - \underbrace{\min\big\{y \,|\, (x,y) \in \mathsf{conv}(X_D)\big\}}_{=\mathsf{vex}_D[f](x)}$$
$$\mathsf{rmcgap}_D[f](x) = \underbrace{\max\big\{y \,|\, (x,y) \in \mathsf{RMC}(X_D)\big\}}_{\stackrel{\text{def}}{=}\mathsf{rmcu}_D[f](x)} - \underbrace{\min\big\{y \,|\, (x,y) \in \mathsf{RMC}(X_D)\big\}}_{\stackrel{\text{def}}{=}\mathsf{rmcl}_D[f](x)}.$$

The relation $\mathsf{rmcgap}_D[f](x) \geq \mathsf{chgap}_D[f](x)$ always holds, and by Corollaries 1 and 2 equality holds when either $\ell = \mathbf{0}$ or $\ell = -u$. We present examples that show that $\mathsf{rmcgap}_D[f](x)$ can be arbitrarily larger than $\mathsf{chgap}_D[f](x)$.

First, let $D_u = [1, u]^3$ for some $u > 1$ and consider the point $\hat{x} = (\frac{u+1}{2}, u, 1)$. The only way $\hat{x}$ can be written as a convex combination of vertices of D_u is $\hat{x} = \frac{1}{2}(1, u, 1) + \frac{1}{2}(u, u, 1)$. Thus, $\mathsf{vex}_{D_u}[f](\hat{x}) = \mathsf{cav}_{D_u}[f](\hat{x})$ so $\mathsf{chgap}_{D_u}[f](\hat{x}) = 0$. Next consider the recursive McCormick relaxation of f over D_u. It is possible to check that $\mathsf{rmcu}_{D_u}[f](\hat{x}) = u^2 + \frac{1-u}{2}$ and $\mathsf{rmcl}_{D_u}[f](\hat{x}) = u + \frac{u-1}{2}$, and therefore

$$\mathsf{rmcgap}_{D_u}[f](\hat{x}) = \mathsf{rmcu}_{D_u}[f](\hat{x}) - \mathsf{rmcl}_{D_u}[f](\hat{x}) = u^2 - u > 0.$$

Since $\mathsf{chgap}_{D_u}[f](\hat{x}) = 0$, this example shows that, if we let $u \to \infty$, the difference in relaxation quality between the convex hull and recursive McCormick relaxations can be arbitrarily large even for fixed $n = 3$.

Now, let $D_n = [-2, 2]^{n-2} \times [0, 2] \times [-2, 2]$, and consider the point $\hat{x} = (2, \ldots, 2, 0, 0, 2)$. Again, the only way $\hat{x}$ can be written as a convex combination of vertices of D_n is $\hat{x} = \frac{1}{2}(2, \ldots, 2, -2, 0, 2) + \frac{1}{2}(2, \ldots, 2, 2, 0, 2)$ and hence $\mathsf{vex}_{D_n}[f](\hat{x}) = \mathsf{cav}_{D_n}[f](\hat{x})$ so $\mathsf{chgap}_{D_n}[f](\hat{x}) = 0$. On the other hand, if we consider the recursive McCormick relaxation of f over D_n, it can be verified that for $n \geq 3$, $\mathsf{rmcu}_{D_n}[f](\hat{x}) = 2^n$ and $\mathsf{rmcl}_{D_n}[f](\hat{x}) = -2^n$ and hence $\mathsf{rmcgap}_{D_n}[f](\hat{x}) = 2^{n+1}$. Thus, by letting $n \to \infty$, we see that even if the bounds ℓ and u do not grow, the difference in relaxation quality between the convex hull and recursive McCormick relaxations can be arbitrarily large.

3 General multilinear functions

We now consider general multilinear functions of the form

$$\phi(x) = \sum_{t \in T} a_t \prod_{j \in J_t} x_j, \tag{6}$$

defined over $x \in [\ell, u]$, and study concave upper bounding and convex lower bounding functions of ϕ over $[\ell, u]$. In Section 3.1 we focus on concave envelopes, and study cases in which the concave envelope of ϕ can be written as a sum of concave envelopes of the individual terms. Many of these results follow from results in [6, 15], but we review them

since they are necessary in what follows. Our main results are in Section 3.2, where we show that for bilinear functions (i.e., $|J_t| \leq 2\ \forall t$), the gap between the McCormick upper and lower bounding functions of ϕ is uniformly within a constant of the gap between the concave and convex envelopes of ϕ.

Recall that the concave and convex envelopes of ϕ have the following representations [19,22]:

$$\mathsf{cav}_{[\ell,u]}[\phi](x) = \max_{\lambda}\Big\{\sum_{k=1}^{2^n}\lambda_k\phi(x^k)\ \Big|\ \sum_{k=1}^{2^n}\lambda_k x^k = x, \lambda \in \Delta_{2^n}\Big\} \tag{7}$$

$$\mathsf{vex}_{[\ell,u]}[\phi](x) = \min_{\lambda}\Big\{\sum_{k=1}^{2^n}\lambda_k\phi(x^k)\ \Big|\ \sum_{k=1}^{2^n}\lambda_k x^k = x, \lambda \in \Delta_{2^n}\Big\} \tag{8}$$

where $x^k, k = 1, \ldots, 2^n$ are the vertices of $[\ell, u]$.

3.1 Concave envelope of a sum of multilinear terms

The first result is an almost immediate consequence of [6] and has been explicitly proved in [15].

Theorem 4 ([15]) *Let $\phi : H \to \mathbb{R}$ be as defined in* (6)*, and assume that $a_t > 0$ for all $t \in T$. Also let $f_t(x) = \prod_{j\in J_t} x_j$ for $t \in T$. Then the concave envelope of ϕ is given by the sum of concave envelopes of f_t:*

$$\mathsf{cav}_H[\phi](x) = \sum_{t\in T} a_t\, \mathsf{cav}_H[f_t](x) \quad \forall x \in H.$$

The condition $a_t > 0$ for all $t \in T$ is necessary, even if ϕ is a bilinear function. Example 2 in Section 3.2.3 provides an example of a bilinear function with a single negative a_t and an $x \in H$ at which the sum of concave envelopes of the individual bilinear terms is strictly larger than the concave envelope of the bilinear function.

Theorem 4 can be generalized to the case $x \in [\ell, u]$, provided $\ell \geq \mathbf{0}$.

Theorem 5 *Let $\ell, u \in \mathbb{R}^n$ satisfy $\mathbf{0} \leq \ell \leq u$ and let $\phi : [\ell, u] \to \mathbb{R}$ be as defined in* (6)*, and assume that $a_t > 0$ for all $t \in T$. Also let $f_t(x) = \prod_{j\in J_t} x_j$ for $t \in T$. Then the concave envelope of ϕ over $[\ell, u]$ is given by the sum of concave envelopes of f_t:*

$$\mathsf{cav}_{[\ell,u]}[\phi](x) = \sum_{t\in T} a_t\, \mathsf{cav}_{[\ell,u]}[f_t](x) \quad \forall x \in [\ell, u].$$

Proof Define $\phi' : H \to \mathbb{R}$ by

$$\begin{aligned}\phi'(x') = \phi\left(\mathrm{Diag}(u-\ell)x' + \ell\right) &= \sum_{t\in T} a_t f_t\left(\mathrm{Diag}(u-\ell)x' + \ell\right)\\ &= \sum_{t\in T} a_t \sum_{k\in K_t} a'_k f'_k(x')\end{aligned}$$

where the functions f'_k have the form $f'_k(x') = \prod_{j \in J_k} x'_j$. Also, $a'_k \geq 0$ since each is a product of ℓ_j and $(u_j - \ell_j)$ terms and $\ell_j \geq 0$. Now, let $\phi'_t(x') = f_t\left(\mathrm{Diag}(u-\ell)x' + \ell\right) = \sum_{k \in K_t} a'_k f'_k(x')$ for $t \in T$. Applying Theorem 4 twice then yields

$$\mathsf{cav}_H[\phi'](x') = \sum_{t \in T} a_t \sum_{k \in K_t} a'_k \,\mathsf{cav}_H[f'_k](x') = \sum_{t \in T} a_t \,\mathsf{cav}_H[\phi'_t](x') \quad \forall x' \in H. \tag{9}$$

Next, because $\phi'_t(x') = f_t(\mathrm{Diag}(u-\ell)x' + \ell)$ and $x \in H$ if and only if $(\mathrm{Diag}(u-\ell)x' + \ell) \in [\ell, u]$, it is not hard to see that

$$\mathsf{cav}_H[\phi'_t](x') = \mathsf{cav}_{[\ell,u]}[f_t](\mathrm{Diag}(u-\ell)x' + \ell), \quad \forall x' \in H. \tag{10}$$

Now, let $x \in [\ell, u]$ and let $x' = \mathrm{Diag}(u-\ell)^{-1}(x-\ell)$ and $y' = \mathsf{cav}_H[\phi'](x')$. Then there exists $\lambda \in \Delta_{2^n}$ such that $\sum_k \lambda_k \tilde{x}^k = x'$ and $\sum_k \lambda_k \phi'(\tilde{x}^k) = y'$, where $\tilde{x}^k, k = 1, \ldots, 2^n$ are the vertices of H. Then, observing that $x^k = \mathrm{Diag}(u-\ell)\tilde{x}^k + \ell$, for $k = 1, \ldots, 2^n$ are the vertices of $[\ell, u]$ we have

$$\sum_{k=1}^{2^n} \lambda_k x^k = \sum_{k=1}^{2^n} \lambda_k \left(\mathrm{Diag}(u-\ell)\tilde{x}^k + \ell\right) = \mathrm{Diag}(u-\ell)x' + \ell = x$$

and so λ is feasible to the linear program (7) defining $\mathsf{cav}_{[\ell,u]}[\phi]$. Also, the objective value of λ in (7) is

$$\sum_{k=1}^{2^n} \lambda_k \phi(x^k) = \sum_{k=1}^{2^n} \lambda_k \phi'(\tilde{x}^k) = y' = \sum_{t \in T} a_t \,\mathsf{cav}_H[\phi'_t](x') = \sum_{t \in T} a_t \,\mathsf{cav}_{[\ell,u]}[f_t](x)$$

where the second-to-last equality follows from (9) and the last equality follows from (10). This proves

$$\mathsf{cav}_{[\ell,u]}[\phi](x) \geq \sum_{t \in T} a_k \,\mathsf{cav}_{[\ell,u]}[f_t](x)$$

and completes the proof as the reverse inequality is immediate. □

The following example shows that for general multilinear functions, the condition $\ell \geq \mathbf{0}$ is necessary.

Example 1 Let $D = [-1, 1] \times [0, 1]^3$ and $\phi(x) = f_1(x) + f_2(x)$ where $f_1(x) = x_1x_2x_3$ and $f_2(x) = x_2x_3x_4$, and consider the point $\hat{x} = (-1, 1/3, 1/3, 1/3)$. For this example, it is easy to verify by solving (7) that $\mathsf{cav}_D[\phi](\hat{x}) = 0$. In addition, (7) can be used to find $\mathsf{cav}_D[f_1](\hat{x}) = 0$ and $\mathsf{cav}_D[f_2](\hat{x}) = 1/3$, and thus $\mathsf{cav}_D[\phi](\hat{x}) < \mathsf{cav}_D[f_1](\hat{x}) + \mathsf{cav}_D[f_2](\hat{x})$.

For bilinear functions, Theorem 4 can be generalized to allow $x \in [\ell, u]$ for any $\ell \leq u$. The arguments are fairly standard, but we provide a proof for completeness.

Corollary 3 *Let $b(x) = \sum_{\{i,j\} \in E} a_{ij} x_i x_j$ for $x \in [\ell, u]$, where $\ell, u \in \mathbb{R}^n$ and E is a set of $\{i, j\}$ pairs, and assume $a_{ij} > 0$ for all $\{i, j\} \in E$. Then the concave envelope of b is equal to the termwise McCormick upper bounding function:*

$$\mathsf{cav}_{[\ell,u]}[b](x) = \sum_{\{i,j\} \in E} a_{ij} \min\{u_j x_i + \ell_i x_j - \ell_i u_j, \ell_j x_i + u_i x_j - u_i \ell_j\} \quad \forall x \in [\ell, u].$$

Proof Define $b' : H \to \mathbb{R}$ by

$$b'(x') = b\left(\mathrm{Diag}(u-\ell)x' + \ell\right) = \sum_{\{i,j\}\in E} a_{ij}\left((u_i-\ell_i)\,x'_i + \ell_i\right)\left(\left(u_j-\ell_j\right)x'_j + \ell_j\right)$$
$$= f'(x') + L(x'),$$

where $f'(x') = \sum_{\{i,j\}\in E} a_{ij}\,(u_i-\ell_i)\left(u_j-\ell_j\right)x'_i x'_j$ is a bilinear function having positive coefficients, and $L(x') = \sum_{\{i,j\}\in E} a_{ij}\left[\ell_j(u_i-\ell_i)x'_i + \ell_i(u_j-\ell_j)x'_j + \ell_i\ell_j\right]$ is an affine function of x'. Thus,

$$\begin{aligned}\mathsf{cav}_H[b'](x') &= \mathsf{cav}_H[f'](x') + L(x')\\ &= \sum_{\{i,j\}\in E} a_{ij}\,(u_i-\ell_i)\left(u_j-\ell_j\right)\min\{x'_i, x'_j\} + L(x')\end{aligned}$$

where the first equation follows because L is an affine function, and the second equation follows from Theorem 4 and from the fact that for $f(x_1,x_2) = x_1x_2$, $\mathsf{cav}_{[0,1]^2}[f](x_1,x_2) = \min\{x_1,x_2\}$. By a simple scaling argument ($x' \in H \Leftrightarrow \mathrm{Diag}(u-\ell)x' + \ell \in [\ell,u]$) it holds that

$$\mathsf{cav}_H[b'](x') = \mathsf{cav}_{[\ell,u]}[b](\mathrm{Diag}(u-\ell)x' + \ell).$$

Now, let $x \in [\ell,u]$ and let $x' = \mathrm{Diag}(u-\ell)^{-1}(x-\ell) \in H$. Then,

$$\begin{aligned}\mathsf{cav}_{[\ell,u]}[b](x) &= \mathsf{cav}_H[b'](x')\\ &= \sum_{\{i,j\}\in E} a_{ij}\,(u_i-\ell_i)\left(u_j-\ell_j\right)\min\{x'_i, x'_j\} + L(x')\\ &= \sum_{\{i,j\}\in E} a_{ij}\min\{u_jx_i + \ell_ix_j - \ell_iu_j,\ \ell_jx_i + u_ix_j - u_i\ell_j\}\end{aligned}$$

where the last equation follows because for each $\{i,j\} \in E$,

$$\begin{aligned}&(u_i-\ell_i)\left(u_j-\ell_j\right)\min\{x'_i,x'_j\} + \ell_j(u_i-\ell_i)x'_i + \ell_i(u_j-\ell_j)x'_j + \ell_i\ell_j\\ &= \min\{(u_j-\ell_j)(x_i-\ell_i), (u_i-\ell_i)(x_j-\ell_j)\} + \ell_j(x_i-\ell_i) + \ell_i(x_j-\ell_j) + \ell_i\ell_j\\ &= \min\{u_jx_i + \ell_ix_j - \ell_iu_j,\ \ell_jx_i + u_ix_j - u_i\ell_j\}.\end{aligned}$$

□

3.2 Approximation results for bilinear functions

In this section, we study the strength of the McCormick relaxation for bilinear functions of the form:

$$b(x) = \sum_{\{i,j\}\in E} a_{ij}x_ix_j \tag{11}$$

for $x \in H$, where E is a subset of unordered pairs of distinct indices in $N = \{1,\ldots,n\}$. Specifically, the McCormick upper bounding function is

$$\mathsf{mcu}_H[b](x) = \max_{(x,y)\in P} \sum_{\{i,j\}\in E} a_{ij}y_{ij}$$

and the McCormick lower bounding function is

$$\mathsf{mcl}_H[b](x) = \min_{(x,y)\in P} \sum_{\{i,j\}\in E} a_{ij}y_{ij}$$

where $P = \{x \in H, y \in [0,1]^{|E|} \mid y_{ij} \geq x_i + x_j - 1,\ y_{ij} \leq x_i,\ y_{ij} \leq x_j,\ \forall\{i,j\} \in E\}$ is the polyhedron obtained by using the McCormick inequalities to bound the bilinear terms x_ix_j.

We are interested in the quality of the McCormick approximation as compared to the relaxation given by the convex and concave envelopes of b. We therefore define

$$\begin{aligned}\mathsf{mcgap}_H[b](x) &= \mathsf{mcu}_H[b](x) - \mathsf{mcl}_H[b](x), \text{ and}\\ \mathsf{chgap}_H[b](x) &= \mathsf{cav}_H[b](x) - \mathsf{vex}_H[b](x).\end{aligned}$$

$\mathsf{mcgap}_H[b](x)$ is a measure of the tightness of the McCormick relaxation of $b(x)$ at each point $x \in H = [0,1]^n$, and likewise for $\mathsf{chgap}_H[b](x)$. In this section, we show that under certain conditions, $\mathsf{mcgap}_H[b](x)$ is uniformly close to $\mathsf{chgap}_H[b](x)$.

We begin in Section 3.2.1 by reviewing some existing results and establishing some new results needed for proving our main theorems. Then, in Section 3.2.2 we give our results for the case $a_{ij} > 0$ for all $\{i,j\} \in E$. In Section 3.2.3 we present our (weaker) results for the general case. Throughout this section we assume $x \in H$. However, all the results can be generalized to $x \in [\ell, u]$ using arguments similar to those in the proof of Corollary 3.

We first introduce some new notation. For a graph $G = (N, E)$, we let $\chi(G)$ be the coloring number of G. Also, when G is associated with weights w_e for $e \in E$, we define $w(E') = \sum_{e\in E'} w_e$ for any $E' \subseteq E$. We also define $E^+ = \{e \in E \mid w_e > 0\}$, $E^- = E \setminus E^+$, and for $E' \subseteq E$, $w^+(E') = \sum_{e\in E^+\cap E'} w_e$ and $w^-(E') = \sum_{e\in E^-\cap E'} w_e$. We let $\mathcal{S} = \{S \mid S \subseteq N\}$ be the set of all subsets of N. For two sets $S_1, S_2 \subseteq N$, $\delta(S_1, S_2) = \{e \in E \mid e \text{ has one end in } S_1 \text{ and one end in } S_2\}$. For any $S \in \mathcal{S}$, we let $\delta(S) = \delta(S, N \setminus S)$ and $\gamma(S) = \{e \in E \mid e \text{ has both ends in } S\}$. Finally, for $i \in N$, we let $\mathcal{S}_i = \{S \in \mathcal{S} \mid i \in S\}$ be the set of subsets that contain element i.

3.2.1 Preliminaries

We first state two existing results that are required for our analysis.

Theorem 6 ([18]) *Let $P = \{x \in H, y \in [0,1]^{|E|} \mid y_{ij} \geq x_i + x_j - 1,\ y_{ij} \leq x_i,\ y_{ij} \leq x_j,\ \forall\{i,j\} \in E\}$. The extreme points of P are all $\{0, 1/2, 1\}$-valued.*

In [18], Theorem 6 is proved for the case that E is the set of edges of a complete graph, but the theorem is also true when E is any subset of edges.

Theorem 7 ([13]) *Consider any graph $G = (N, E)$ having $|N|$ even and weights w_e for $e \in E$. There exists a matching $M \subseteq E$, with*

$$w(M) \geq \frac{w(E)}{|N| - 1}.$$

The following corollary is a slight strengthening of the simple result that there exists a cut with weight at least half the weight of all edges in the graph (see, e.g., Theorem 5.1 in [16]). It is a slight improvement on a result in [4]. The slight improvement is important for our results and can be obtained using arguments from [10] using Theorem 7 in place of the (weaker) bound on the size of a matching used in [4]. (See also the discussion in [12]).

Corollary 4 *Let $G = (N, E)$ be a graph with $|N|$ even and weights w_e for $e \in E$. Then there exist cuts $C_1, C_2 \subseteq E$ in G having*

$$w(C_1) \geq \frac{1}{2}w(E) + \frac{\sum_{e\in E}|w_e|}{2(|N|-1)}, \tag{12}$$

$$w(C_2) \leq \frac{1}{2}w(E) - \frac{\sum_{e\in E}|w_e|}{2(|N|-1)}. \tag{13}$$

Proof By applying Theorem 7 using weights $w'_e = |w_e|$, there exists a matching M in the graph (N, E) with $\sum_{e\in M}|w_e| \geq \sum_{e\in E}|w_e|/(|N|-1)$. We construct a random cut $\tilde{C}$ to be defined by the edges between the sets S and $N \setminus S$ which are generated as follows. For every edge $e = \{i, j\} \in M$, if $w_e > 0$ we assign i to S and j to $N \setminus S$ with probability $1/2$ and assign j to S and i to $N \setminus S$ with probability $1/2$; if $w_e \leq 0$ we assign i and j to S with probability $1/2$ and assign i and j to $N \setminus S$ with probability $1/2$. Thus, with probability 1, every positive weight edge in M is in the cut $\tilde{C}$ and every nonpositive weight edge in M is not in the cut $\tilde{C}$, but every node that was matched by an edge in M has equal probability of being in S or $N \setminus S$. For every node i that was not matched by M, we assign i to S with probability $1/2$ and to $N \setminus S$ with probability $1/2$. Thus, any edge $e \in E \setminus M$ has probability $1/2$ of being in the cut $\tilde{C}$. Therefore, the expected weight of the cut is:

$$\begin{aligned} E[w(\tilde{C})] &= w^+(M) + \frac{1}{2}\sum_{e\in E\setminus M} w_e = w^+(M) + \frac{1}{2}(w(E) - w^+(M) - w^-(M)) \\ &= \frac{1}{2}w(E) + \frac{1}{2}(w^+(M) - w^-(M)) \\ &= \frac{1}{2}w(E) + \frac{1}{2}\sum_{e\in M}|w_e| \geq \frac{1}{2}w(E) + \frac{\sum_{e\in E}|w_e|}{2(|N|-1)}. \end{aligned}$$

This implies there exists a cut C_1 that achieves at least the value of the expected weight of this random cut, proving (12).

Existence of a cut C_2 satisfying (13) is established with a nearly identical argument as for C_1, with the exception being that given a matching M with $\sum_{e\in M}|w_e| \geq \sum_{e\in E}|w_e|/(|N|-1)$, a random cut $\tilde{C}$ is constructed by placing i and j in the same node set (S or $N \setminus S$ with equal probability) if $w_{\{i,j\}} > 0$ and by placing i and j in different node sets if $w_{\{i,j\}} \leq 0$. □

This result can be strengthened further for graphs that have a small coloring number when all weights are nonnegative.

Lemma 1 *Let $G = (N, E)$ be a graph with $\chi(G)$ even, and weights $w_e \geq 0$ for $e \in E$. Then there exist a cut C in G with*

$$w(C) \geq \frac{1}{2}w(E) + \frac{1}{2(\chi(G)-1)}w(E),$$

Proof Let $\chi = \chi(G)$ and let $S_1, \ldots, S_\chi$ be a partition of N such that $\gamma(S_i) = \emptyset$ for all $i = 1, \ldots, k$. (I.e., these sets define a coloring of size χ.) Define a complete graph G' with vertices $N' = \{1, \ldots, \chi\}$, and define $\bar{w}_{ij} = w(\delta(S_i, S_j))$ for $1 \leq i < j \leq \chi$ as the weights

on the edges, E', in G'. By definition, $\bar{w}(E') = w(E)$. Applying Corollary 4 to the graph G', there exists a cut C' in G' with

$$\bar{w}(C') \geq \frac{1}{2}\bar{w}(E') + \frac{1}{2(\chi-1)}\bar{w}(E') = \frac{1}{2}w(E) + \frac{1}{2(\chi-1)}w(E).$$

Now let C be the set of edges in E defined by $C = \bigcup_{\{i,j\}\in C'} \delta(S_i, S_j)$. Since $w(C) = \bar{w}(C')$ and C is a cut in G, this proves the result. □

Due to Theorem 6, vectors x that are $\{0, 1/2, 1\}$-valued play an important role in our analysis. We therefore determine $\mathsf{mcgap}_H[b](x)$ and find bounds on $\mathsf{cav}_H[b](x)$ and $\mathsf{vex}_H[b](x)$ for such vectors.

Lemma 2 *Let $x \in \mathbb{R}^n$ be $\{0, 1/2, 1\}$-valued and let $T_1 = \{i \in N \mid x_i = 1\}$ and $T_f = \{i \in N \mid x_i = 1/2\}$. Then*

$$\mathsf{mcgap}_H[b](x) = \tfrac{1}{2} \sum_{\{i,j\}\in\gamma(T_f)} |a_{ij}|.$$

Proof We first derive an expression for $\mathsf{mcgap}_H[b](x)$ for any $x \in H$:

$$\mathsf{mcgap}_H[b](x) = \sum_{\{i,j\}\in E} |a_{ij}| \left(\min\{x_i, x_j\} - \max\{x_i + x_j - 1, 0\}\right). \tag{14}$$

Indeed,

$$\begin{aligned}
\mathsf{mcgap}_H[b](x) &= \mathsf{mcu}_H[b](x) - \mathsf{mcl}_H[b](x) \\
&= \sum_{\{i,j\}\in E^+} a_{ij} \min\{x_i, x_j\} + \sum_{\{i,j\}\in E^-} a_{ij} \max\{x_i + x_j - 1, 0\} \\
&\quad - \left(\sum_{\{i,j\}\in E^+} a_{ij} \max\{x_i + x_j - 1, 0\} + \sum_{\{i,j\}\in E^-} a_{ij} \min\{x_i, x_j\} \right) \\
&= \sum_{\{i,j\}\in E} |a_{ij}|(\min\{x_i, x_j\} - \max\{x_i + x_j - 1, 0\})
\end{aligned}$$

Now, if $\{i,j\} \in \gamma(T_1)$, and hence $i, j \in T_1$, then $\min\{x_i, x_j\} = \max\{x_i + x_j - 1, 0\} = 1$. If $\{i,j\} \in \delta(T_1, T_f)$, then $\min\{x_i, x_j\} = \max\{x_i + x_j - 1, 0\} = 1/2$. If $\{i,j\} \in \gamma(T_f)$, then $x_i = x_j = 1/2$ and hence $\min\{x_i, x_j\} = 1/2$ and $\max\{x_i + x_j - 1, 0\} = 0$. Finally, in all other cases for $\{i,j\}$, $\min\{x_i, x_j\} = \max\{x_i + x_j - 1, 0\} = 0$. Thus, the result follows from (14). □

Lemma 3 *Let $x \in \mathbb{R}^n$ be $\{0, 1/2, 1\}$-valued and let $T_1 = \{i \in N \mid x_i = 1\}$ and $T_f = \{i \in N \mid x_i = 1/2\}$.*

(a) If $a_{ij} \geq 0$ for all $\{i,j\} \in E$, then

$$\mathsf{vex}_H[b](x) \leq a(\gamma(T_1)) + \frac{1}{2}a(\delta(T_1, T_f)) + \frac{1}{4}a(\gamma(T_f)) - \frac{1}{4(\chi(G)-1)}a(\gamma(T_f)). \tag{15}$$

(b) If the weights $a_{ij}, \{i,j\} \in E$ have mixed-sign, then

$$\mathsf{vex}_H[b](x) \leq a(\gamma(T_1)) + \frac{1}{2}a(\delta(T_1,T_f)) + \frac{1}{4}a(\gamma(T_f)) - \frac{\sum_{\{i,j\}\in\gamma(T_f)}|a_{ij}|}{4(|N|-1)} \quad (16)$$

and

$$\mathsf{cav}_H[b](x) \geq a(\gamma(T_1)) + \frac{1}{2}a(\delta(T_1,T_f)) + \frac{1}{4}a(\gamma(T_f)) + \frac{\sum_{\{i,j\}\in\gamma(T_f)}|a_{ij}|}{4(|N|-1)}. \quad (17)$$

Proof First, observe that for every vertex x^k of H, if we let $S_k = \{i \mid x_i^k = 1\}$ then $b(x^k) = \sum_{\{i,j\}\in E} a_{ij}x_i^k x_j^k = \sum_{\{i,j\}\in\gamma(S_k)} a_{ij} = a(\gamma(S_k))$. Thus, we can rewrite the LP (8) defining $\mathsf{vex}_H[b](x)$ as follows:

$$\mathsf{vex}_H[b](x) = \min_{\lambda\in\Delta_{2^n}} \sum_{S\in\mathcal{S}} a(\gamma(S))\lambda_S \quad (18a)$$

$$\text{s.t.} \sum_{S\in\mathcal{S}_i} \lambda_S = x_i, \quad i = 1,\ldots,n. \quad (18b)$$

Now, let $C = \delta(U_1,U_2)$ be a maximum weight cut in the subgraph G_f of G induced by the nodes T_f, where U_1 and U_2 are the node sets defining the cut ($U_1 \cup U_2 = T_f$ and $U_1 \cap U_2 = \emptyset$). Let $S_1 = T_1 \cup U_1$ and $S_2 = T_1 \cup U_2$, and construct a solution to (18) by letting $\lambda_{S_1} = \lambda_{S_2} = 1/2$, and $\lambda_S = 0$ otherwise. Clearly, $\lambda \in \Delta_{2^n}$. Also, if $i \in T_1$ then $i \in S_1 \cap S_2$, so $\sum_{S\in\mathcal{S}_i}\lambda_S = \lambda_{S_1} + \lambda_{S_2} = 1 = x_i$. If $i \in T_f$, then i is in either S_1 or S_2, so $\sum_{S\in\mathcal{S}_i}\lambda_S = 1/2 = x_i$. Otherwise, i is in neither S_1 nor S_2, and hence (18b) is satisfied as well. Thus, because λ is one feasible solution to (18),

$$\mathsf{vex}_H[b](x) \leq \frac{1}{2}\big(a(\gamma(S_1)) + a(\gamma(S_2))\big). \quad (19)$$

Next, using the definitions of S_1 and S_2, we observe that for $i = 1,2$

$$a(\gamma(S_i)) = a(\gamma(U_i)) + a(\delta(T_1,U_i)) + a(\gamma(T_1)).$$

Then, observing that $a(\delta(T_1,U_1))+a(\delta(T_1,U_2)) = a(\delta(T_1,T_f))$ and $a(\gamma(U_1))+a(\gamma(U_2)) = a(\gamma(T_f)) - a(\delta(U_1,U_2)) = a(\gamma(T_f)) - a(C)$ yields

$$a(\gamma(S_1)) + a(\gamma(S_2)) \qquad = 2a(\gamma(T_1)) + a(\delta(T_1,T_f)) + a(\gamma(T_f)) - a(C). \quad (20)$$

Now, if $a_{ij} \geq 0$ for all $\{i,j\} \in E$, then because the coloring number of G_f is no larger than the coloring number of G, Lemma 1 implies

$$a(C) \geq \frac{1}{2}a(\gamma(T_f)) + \frac{1}{2(\chi(G)-1)}a(\gamma(T_f)).$$

Combining this with (20) and (19) yields part (a). When the weights a_{ij} aren't necessarily nonnegative, inequality (12) of Lemma 4 yields

$$a(C) \geq \frac{1}{2}a(\gamma(T_f)) + \frac{\sum_{\{i,j\}\in E}|a_{ij}|}{2(|N|-1)},$$

which, combined with (20) and (19), proves (16) for part (b).

The proof of (17) is similar to that of (16), except that we use inequality (13) in Lemma 4 to obtain a cut C_2 such that

$$a(C_2) \leq \frac{1}{2}a(\gamma(T_f)) - \frac{\sum_{\{i,j\}\in E}|a_{ij}|}{2(|N|-1)}.$$

This cut can then be used to construct a feasible solution to the maximization problem defining $\mathsf{cav}_H[b](x)$ with objective value equal to the lower bound in (17). □

3.2.2 Bilinear functions with positive weights

In this section, we consider bilinear functions having *positive* weights: $a_{ij} > 0$ for all $\{i,j\} \in E$. We first state the main result.

Theorem 8 *Let $G = (N, E)$ have a coloring of size χ, and let $b(x)$ be a bilinear function of the form* (11) *with $a_{ij} > 0$ for all $\{i,j\} \in E$. Then if χ is even,*

$$\mathsf{mcgap}_H[b](x) \leq \left(2 - \frac{2}{\chi}\right) \mathsf{chgap}_H[b](x) \quad \forall x \in H,$$

and if χ is odd,

$$\mathsf{mcgap}_H[b](x) \leq \left(2 - \frac{2}{\chi+1}\right) \mathsf{chgap}_H[b](x) \quad \forall x \in H.$$

Note that the theorem implies the result that for bipartite graphs (graphs with coloring of size two) the McCormick envelopes provide the convex lower and upper envelopes, which was first proved in [5,9].

Proof We prove the case where χ is even. The case where χ is odd is an immediate consequence since if the coloring number $\chi(G)$ of a graph is odd, then it has an even coloring of size $\chi(G)+1$. Let $K = 2 - \frac{2}{\chi}$. We need to prove

$$\min_{x \in H} \big(K\, \mathsf{chgap}_H[b](x) - \mathsf{mcgap}_H[b](x)\big) \geq 0. \tag{21}$$

Next, because $a_{ij} > 0$ for all $\{i,j\} \in E$, Theorem 4 applies and hence $\mathsf{cav}_H[b](x) = \mathsf{mcu}_H[b](x)$. Using this, the definitions of $\mathsf{chgap}_H[b]$ and $\mathsf{mcgap}_H[b]$, and expanding the definition of $\mathsf{mcl}_H[b](x)$, the minimization problem in (21) is equivalent to:

$$\min_{(x,y) \in P} \Big((K-1)\, \mathsf{cav}_H[b](x) - K\, \mathsf{vex}_H[b](x) + \sum_{\{i,j\} \in E} a_{ij} y_{ij}\Big)$$

where $P = \{x \in H, y \in [0,1]^{|E|} \mid y_{ij} \geq x_i + x_j - 1,\ y_{ij} \leq x_i,\ y_{ij} \leq x_j,\ \forall \{i,j\} \in E\}$ is as defined in Theorem 6. Then, because $\mathsf{cav}_H[b](x)$ and $-\mathsf{vex}_H[b](x)$ are concave functions, the above problem is a concave minimization problem over a polytope, and hence achieves its minimum at an extreme point. Theorem 6 then implies that it is sufficient to prove

$$K\, \mathsf{chgap}_H[b](x) - \mathsf{mcgap}_H[b](x) \geq 0 \tag{22}$$

for all $\{0, 1/2, 1\}$ vectors x.

Therefore, let x be an arbitrary $\{0,1/2,1\}$-valued vector, and let $T_1 = \{i \in N \mid x_i = 1\}$ and $T_f = \{i \in N \mid x_i = 1/2\}$. Since $a_{ij} > 0$ for all $\{i,j\} \in E$, Lemma 2 then implies

$$\mathsf{mcgap}_H[b](x) = \frac{1}{2} \sum_{\{i,j\} \in \gamma(T_f)} |a_{ij}| = \frac{1}{2} a(\gamma(T_f)). \tag{23}$$

Next, again using Theorem 4,

$$\begin{aligned} \mathsf{cav}_H[b](x) = \mathsf{mcu}_H[b](x) &= \sum_{\{i,j\} \in E} a_{ij} \min\{x_i, x_j\} \\ &= a(\gamma(T_1)) + \frac{1}{2} a(\delta(T_1, T_f)) + \frac{1}{2} a(\gamma(T_f)), \end{aligned}$$

where the last equality follows because $\min\{x_i, x_j\} = 1$ for $\{i,j\} \in \gamma(T_1)$, $\min\{x_i, x_j\} = 1/2$ for $\{i,j\} \in \gamma(T_f) \cup \delta(T_1, T_f)$, and $\min\{x_i, x_j\} = 0$ otherwise. Combining this with (15) from Lemma 3 and (23) yields

$$\begin{aligned}\mathsf{chgap}_H[b](x) = \mathsf{cav}_H[b](x) - \mathsf{vex}_H[b](x) &\geq \frac{1}{4}\left(1 + \frac{1}{\chi - 1}\right) a(\gamma(T_f)) \\ &= \frac{\chi}{2(\chi-1)}\,\mathsf{mcgap}_H[b](x).\end{aligned}$$

Rearranging yields

$$\mathsf{mcgap}_H[b](x) \leq \frac{2(\chi-1)}{\chi}\,\mathsf{chgap}_H[b](x) = \left(2 - \frac{2}{\chi - 1}\right)\mathsf{chgap}_H[b](x)$$

and so indeed (22) holds. □

3.2.3 General bilinear functions

In this section, we consider bilinear functions that may have both positive and negative coefficients on the bilinear terms. We first state the main result.

Theorem 9 *Let $G = (N, E)$ and let $b(x)$ be a bilinear function of the form* (11) *over $x \in H$. Then if $|N|$ is even,*

$$\mathsf{mcgap}_H[b](x) \leq (|N| - 1)\,\mathsf{chgap}_H[b](x) \quad \forall x \in H, \tag{24}$$

and if $|N|$ is odd,

$$\mathsf{mcgap}_H[b](x) \leq |N|\,\mathsf{chgap}_H[b](x) \quad \forall x \in H.$$

Proof As in the proof of Theorem 8, we restrict attention to the case where $|N|$ is even. First, for $x_i, x_j \in [0,1]$ observe that

$$\begin{aligned}\min\{x_i, x_j\} - \max\{x_i + x_j - 1, 0\} &= \min\{x_i, x_j\} + \min\{1 - x_i - x_j, 0\} \\ &= \min\big\{x_i + \min\{1 - x_i - x_j, 0\}, x_j + \min\{1 - x_i - x_j, 0\}\big\} \\ &= \min\{x_i, x_j, 1 - x_i, 1 - x_j\}.\end{aligned}$$

Thus, using this in (14) we can write $\mathsf{mcgap}_H[b](x)$ as

$$\begin{aligned}\mathsf{mcgap}_H[b](x) &= \sum_{\{i,j\} \in E} |a_{ij}| \min\{x_i, x_j, 1 - x_i, 1 - x_j\} \\ &= \max_{(x,z) \in Q} \sum_{\{i,j\} \in E} |a_{ij}| z_{ij}\end{aligned}$$

where $Q = \{x \in H, z \in \mathbb{R}^{|E|} \mid z_{ij} + x_i \leq 1, z_{ij} + x_j \leq 1, z_{ij} \leq x_i, z_{ij} \leq x_j, \forall \{i,j\} \in E\}$. All the constraints of Q are of the form $z_{ij} - x_i \leq 0$ or $z_{ij} + x_i \leq 1$, and hence have the form of the constraint matrix of a 2-SAT problem. Thus, the results of [11] imply that all vertices of Q are $\{0, 1/2, 1\}$-valued.

Now, we need to prove

$$\min_{x \in H}\big((|N| - 1)\,\mathsf{chgap}_H[b](x) - \mathsf{mcgap}_H[b](x)\big) \geq 0.$$

This minimization problem is equivalent to:

$$\min_{(x,z)\in Q}\Big((|N|-1)\,\mathsf{chgap}_H[b](x) - \sum_{\{i,j\}\in E}|a_{ij}|z_{ij}\Big)$$

Since $\mathsf{chgap}_H[b](x)$ is a concave function of x, this is a concave minimization problem over the polyhedron Q, and hence has an extreme point optimal solution. Thus, just as in the proof of Theorem 8, it is sufficient to show that (24) holds for $\{0, 1/2, 1\}$-valued x.

Thus, let x be any $\{0, 1/2, 1\}$-valued vector. Using (16) and (17) from Lemma 3 to bound both $\mathsf{vex}_H[b](x)$ and $\mathsf{cav}_H[b](x)$ yields

$$\begin{aligned}\mathsf{chgap}_H[b](x) &= \mathsf{cav}_H[b](x) - \mathsf{vex}_H[b](x)\\ &\geq \frac{1}{4(|N|-1)}\Big(\sum_{\{i,j\}\in\gamma(T_f)}|a_{ij}| + \sum_{\{i,j\}\in\gamma(T_f)}|a_{ij}|\Big)\\ &= \frac{1}{(|N|-1)}\,\mathsf{mcgap}_H[g](x)\end{aligned}$$

by Lemma 2, completing the proof. □

The bound in Theorem 9 is significantly weaker than Theorem 8 which provides a constant approximation guarantee; in this case, the approximation factor is n. In §4 we present numerical examples that suggest this bound is not tight, and we leave it as an open question whether there is a constant factor approximation. The following example shows that even for bipartite graphs, when the weights have mixed signs it is possible that $\mathsf{chgap}_H[b](x) < \mathsf{mcgap}_H[b](x)$, which is in contrast to the case when the weights are all nonnegative.

Example 2 Consider the bipartite graph with $n = 4$ nodes and edge set $E = \{(1,3),(1,4),(2,3),(2,4)\}$ with weights $a_{14} = -1$ and $a_{ij} = 1$ otherwise, and consider the point $x = (1/2, 1/2, 1/2, 1/2)$. Then $\mathsf{mcgap}_H[b](x) = (1/2)\sum_{\{i,j\}\in E}|a_{ij}| = 2$. For $\mathsf{cav}_H[b](x)$, the optimal value sets $\lambda_{\{1,3\}} = \lambda_{\{2,4\}} = 1/2$ and achieves value $(1/2)(a_{13} + a_{24}) = 1$ and for $\mathsf{vex}_H[b](x)$ the optimal value sets $\lambda_{\{1,4\}} = \lambda_{\{2,3\}} = 1/2$ and achieves the value $(1/2)(a_{14} + a_{23}) = 0$. Thus, $\mathsf{chgap}_H[b](x) = 1 < 2 = \mathsf{mcgap}_H[b](x)$. Note also that $\mathsf{mcu}_H[b](x) = 3/2 > 1 = \mathsf{cav}_H[b](x)$, showing the necessity of $a_t > 0$ in Theorem 4, even for the case of a bilinear function in which $G = (N, E)$ is bipartite.

Theorem 8 provides a worst-case approximation guarantee for bilinear functions having nonnegative weights that increases with the coloring number of the graph underlying a bilinear function. Since graphs with small coloring number tend to be less dense, this suggests that the McCormick relaxation gap will generally be closer to the convex hull relaxation gap for sparser graphs. The next result provides further support for this intuition, regardless of the signs of the edge weights. Given a graph $G = (N, E)$ and weights a_{ij} for $\{i, j\} \in E$, for any $E' \subseteq E$ we denote $b_{E'}$ as the bilinear function using only the terms in E':

$$b_{E'}(x) = \sum_{\{i,j\}\in E'} a_{ij}x_ix_j.$$

Theorem 10 *Let $E' \subseteq E$. Then, for any $x \in H$,*

$$\mathsf{mcgap}_H[b_{E'}](x) - \mathsf{chgap}_H[b_{E'}](x) \leq \mathsf{mcgap}_H[b_E](x) - \mathsf{chgap}_H[b_E](x).$$

Proof We prove the equivalent inequality:

$$\mathsf{mcgap}_H[b_E](x) - \mathsf{mcgap}_H[b_{E'}](x) \geq \mathsf{chgap}_H[b_E](x) - \mathsf{chgap}_H[b_{E'}](x). \tag{25}$$

We prove the result holds for $E' = E \setminus \{k,l\}$ where $\{k,l\}$ is an arbitrary edge in E, which implies the result for any $E' \subseteq E$ by induction.

First suppose $a_{kl} > 0$. Then, $\mathsf{mcu}_H[b_E](x) - \mathsf{mcu}_H[b_{E'}](x) = a_{kl}\max\{x_k + x_l - 1, 0\}$ and $\mathsf{mcl}_H[b_E](x) - \mathsf{mcl}_H[b_{E'}](x) = a_{kl}\min\{x_k, x_l\}$. Hence, $\mathsf{mcgap}_H[b_E](x) - \mathsf{mcgap}_H[b_{E'}](x) = a_{kl}\big(\max\{x_k + x_l - 1, 0\} - \min\{x_k, x_l\}\big)$. Similarly, if $a_{kl} < 0$, then $\mathsf{mcgap}_H[b_E](x) - \mathsf{mcgap}_H[b_{E'}](x) = -a_{kl}\big(\max\{x_k + x_l - 1, 0\} - \min\{x_k, x_l\}\big)$. Thus, for any a_{kl},

$$\mathsf{mcgap}_H[b_E](x) - \mathsf{mcgap}_H[b_{E'}](x) = |a_{kl}|\left(\max\{x_k + x_l - 1, 0\} - \min\{x_k, x_l\}\right). \tag{26}$$

Now, suppose again $a_{kl} > 0$ and consider the linear program defining $\mathsf{cav}_H[b_E](x)$:

$$\mathsf{cav}_H[b_E](x) = \max_{\lambda \in \Delta_{2^n}} \sum_{S \in \mathcal{S}} a(\gamma^E(S))\lambda_S \tag{27a}$$

$$\text{s.t.} \sum_{S \in \mathcal{S}_i} \lambda_S = x_i, \quad i = 1, \ldots, n \tag{27b}$$

where we have made the dependence on the edge set E explicit: $\gamma^E(S) = \{\{i,j\} \in E \mid i \in S, j \in S\}$. Let λ^E be an optimal solution to (27). Clearly, λ^E is also a feasible solution to the problem (27) when E' replaces E. Thus,

$$\begin{aligned}
\mathsf{cav}_H[b_E](x) - \mathsf{cav}_H[b_{E'}](x) &\leq \sum_{S \in \mathcal{S}} a(\gamma^E(S))\lambda^E_S - \sum_{S \in \mathcal{S}} a(\gamma^{E'}(S))\Big)\lambda^E_S \\
&= \sum_{S \in \mathcal{S}: \{k,l\} \in \gamma^E(S)} \lambda^E_S \left(a(\gamma^E(S)) - a(\gamma^{E'}(S))\right) \\
&= \sum_{S \in \mathcal{S}_k \cap \mathcal{S}_l} a_{kl}\lambda^E_S.
\end{aligned}$$

But, (27b) implies $\sum_{S \in \mathcal{S}_k \cap \mathcal{S}_l} \lambda^E_S \leq x_k$ and $\sum_{S \in \mathcal{S}_k \cap \mathcal{S}_l} \lambda^E_S \leq x_l$ and hence,

$$\mathsf{cav}_H[b_E](x) - \mathsf{cav}_H[b_{E'}](x) \leq a_{kl}\min\{x_k, x_l\}. \tag{28}$$

Now let λ^E be an optimal solution to the linear program defining $\mathsf{vex}_H[b_E](x)$, which is (27) with $\max$ replaced by $\min$. As λ^E is also feasible to the LP defining $\mathsf{vex}_H[b_{E'}](x)$, we have, similar to the argument for cav_H,

$$\begin{aligned}
\mathsf{vex}_H[b_E](x) - \mathsf{vex}_H[b_{E'}](x) &\geq \sum_{S \in \mathcal{S}} a(\gamma^E(S))\lambda^E_S - \sum_{S \in \mathcal{S}} a(\gamma^{E'}(S))\Big)\lambda^E_S \\
&= \sum_{S \in \mathcal{S}_k \cap \mathcal{S}_l} a_{kl}\lambda^E_S.
\end{aligned}$$

Next, (27b) implies

$$x_k + x_l = \sum_{S \in \mathcal{S}_k} \lambda^E_S + \sum_{S \in \mathcal{S}_l} \lambda^E_S = \sum_{S \in \mathcal{S}_k \cup \mathcal{S}_l} \lambda^E_S + \sum_{S \in \mathcal{S}_k \cap \mathcal{S}_l} \lambda^E_S \leq 1 + \sum_{S \in \mathcal{S}_k \cap \mathcal{S}_l} \lambda^E_S.$$

Since also $\lambda_S^E \geq 0$ this implies

$$\mathsf{vex}_H[b_E](x) - \mathsf{vex}_H[b_{E'}](x) \geq \sum_{S \in \mathcal{S}_k \cap \mathcal{S}_l} a_{kl} \lambda_S^E \geq a_{kl} \max\{x_k + x_l - 1, 0\}.$$

Combining this with (28) implies

$$\begin{aligned}
&\mathsf{chgap}_H[b_E](x) - \mathsf{chgap}_H[b_{E'}](x) \\
&\qquad = \mathsf{cav}_H[b_E](x) - \mathsf{vex}_H[b_E](x) - \Big(\mathsf{cav}_H[b_{E'}](x) - \mathsf{vex}_H[b_{E'}](x)\Big) \\
&\qquad \leq a_{kl} \left(\max\{x_k + x_l - 1, 0\} + \min\{x_k, x_l\}\right) \\
&\qquad = \mathsf{mcgap}_H[b_E](x) - \mathsf{mcgap}_H[b_{E'}](x).
\end{aligned}$$

The argument for $a_{kl} < 0$ is similar, with the only difference being that the inequality $\sum_{S \in \mathcal{S}_k \cap \mathcal{S}_l} \lambda_S^E \leq \min\{x_k, x_l\}$ is needed to bound $\mathsf{vex}_H[b_E](x) - \mathsf{vex}_H[b_{E'}](x)$ and the inequality $\sum_{S \in \mathcal{S}_k \cap \mathcal{S}_l} \lambda_S^E \geq \max\{x_k + x_l - 1, 0\}$ is needed to bound $\mathsf{cav}_H[b_E](x) - \mathsf{cav}_H[b_{E'}](x)$.

□

4 Numerical experiments

In this section we present some numerical examples that illustrate and complement the theory we presented in the previous sections.

First we look at some experiments related to the approximation results for bilinear functions. We are interested in understanding how tight our results are for both the positive coefficients case (Theorem 8) and the mixed-sign coefficients case (Theorem 9). Also, inspired by Theorem 10, we are interested in the effect the graph density has on the quality of the McCormick relaxation compared to the convex hull relaxation.

In our first experiment, we fixed the dimension at $n = 7$ and randomly generated 4000 graphs with varying density. We consider two cases for the coefficients on the bilinear terms appearing in the corresponding bilinear function: (1) all coefficients are positive one, and (2) coefficients have mixed-sign, having '+1' with probability $3/4$ and '-1' with probability $1/4$. For each random graph, we computed the maximum ratio between the McCormick relaxation gap and the convex hull relaxation gap of the corresponding bilinear function. Specifically, we calculated: $\max_{x \in H} \{\mathsf{mcgap}_H[b](x) / \mathsf{chgap}_H[b](x)\}$. This maximum was found by calculating $\mathsf{mcgap}_H[b](x)$ and $\mathsf{chgap}_H[b](x)$ for all 3^7 $\{0, 1/2, 1\}$-valued points in H, where the linear programs (7) and (8) were used to calculate $\mathsf{chgap}_H[b](x)$ for each of these points.

Table 1 displays the results summarized by coloring number. For each coloring number from two to seven, we report the average, maximum, and mode of the maximum ratio taken over all graphs that had that coloring number. For the mode, we also report the percentage of the graphs that achieved that quantity. These results show that the bound of Theorem 8 is tight for coloring number up to seven. Also, the vast majority of the randomly generated graphs achieved this worst-case bound. In contrast, when the coefficients have mixed-sign, the bound of Theorem 9 does not appear tight. The maximum observed ratio was three, in contrast to the bound of $|N| = 7$ given by the theorem. In addition, the bound in Theorem 9 does not depend on the coloring number, but these results show that the worst-case ratio does tend to increase with coloring number.

	Positive Coefficients			Mixed-Sign Coefficients		
	Max Ratio			Max Ratio		
χ	avg	max	mode(%)	avg	max	mode(%)
2	1.000	1.000	1.000(100)	1.111	2.000	1.000(88.7)
3	1.487	1.500	1.500(94.9)	1.706	2.250	1.500(41.5)
4	1.500	1.500	1.500(100)	1.902	2.500	2.000(63.0)
5	1.667	1.667	1.667(99.8)	2.051	2.600	2.000(41.4)
6	1.667	1.667	1.667(100)	2.205	3.000	2.500(54.1)
7	1.750	1.750	1.750(100)	2.294	3.000	2.500(61.4)

Table 1 Maximum gap ratio for random graphs of size 7, summarized by coloring number.

We also summarized our results by graph density in Table 2. The average, maximum, and mode of the worst-case ratios is uniformly increasing as the graph density increases. These results reinforce the intuition provided by Theorem 10 that the difference between the McCormick relaxation and the convex hull relaxation is more significant for denser graphs.

	Positive Coefficients			Mixed-Sign Coefficients		
	Max Ratio			Max Ratio		
density	avg	max	mode(%)	avg	max	mode(%)
0.0–0.1	0.000	0.000	0.000(100)	0.000	0.000	0.000 (100)
0.1–0.2	1.000	1.000	1.000(100)	1.000	1.000	1.000 (100)
0.2–0.3	1.049	1.500	1.000(90.1)	1.090	2.000	1.000 (83.5)
0.3–0.4	1.365	1.500	1.500(67.6)	1.544	2.000	1.500 (55.3)
0.4–0.5	1.494	1.500	1.500(98.4)	1.758	2.250	2.000 (41.1)
0.5–0.6	1.499	1.667	1.500(99.5)	1.859	2.250	2.000 (57.5)
0.6–0.7	1.507	1.667	1.500(95.8)	1.918	2.500	2.000 (86.2)
0.7–0.8	1.542	1.667	1.500(74.9)	1.970	2.500	2.000 (63.1)
0.8–0.9	1.637	1.667	1.667(81.9)	2.032	3.000	2.000 (51.7)
0.9–1.0	1.717	1.750	1.750(60.1)	2.264	3.000	2.500 (57.5)

Table 2 Maximum gap ratio for random graphs of size 7, summarized by density.

We next consider multilinear functions having terms with more than two variables defined over $[\ell, u]$. We conducted some numerical experiments to see how the convex hull relaxation compares to two weaker relaxations: (1) the *recursive McCormick* relaxation, obtained by independently applying recursive McCormick to each of the terms, and (2) the *term-by-term* relaxation, obtained by using the concave and convex envelopes of each of the terms. For these computations, we again used the linear programs (7) and (8) to calculate $\mathsf{chgap}_H[\phi](x)$ for a given point x. The term-by-term relaxation was calculated by using the formulation of (7) and (8) for each product term *independently*. Corollary 5 states that if $\ell \geq 0$ and the coefficients on all terms are positive, the concave upper bounding function given by the term-by-term relaxation is equal to the concave envelope. We are interested in seeing how the recursive McCormick and term-by-term relaxations perform more generally. As an example, we consider the following function:

$$\phi(x) = x_1x_2x_3x_4x_5 + x_1x_2x_3x_4 + x_1x_3x_4x_5 + x_2x_3x_5 + x_1x_3x_5 + x_4x_5 + x_1x_2,$$

which has multiple terms of different sizes, all with positive coefficients. We compare the term-by-term relaxation and recursive McCormick relaxations to the convex hull relaxation of the function over two different domains: $[1, 2]^5$ and $[-1, 2]^5$. Figure 2, for the $[1, 2]^5$ case,

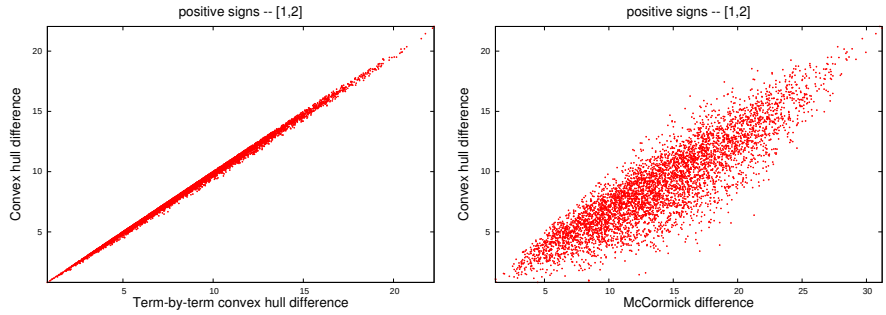

Fig. 2 Scatter plots comparing the term-by-term (left) and recursive McCormick (right) relaxation gaps to the convex hull relaxation gap for the function ϕ defined over $[1,2]^5$.

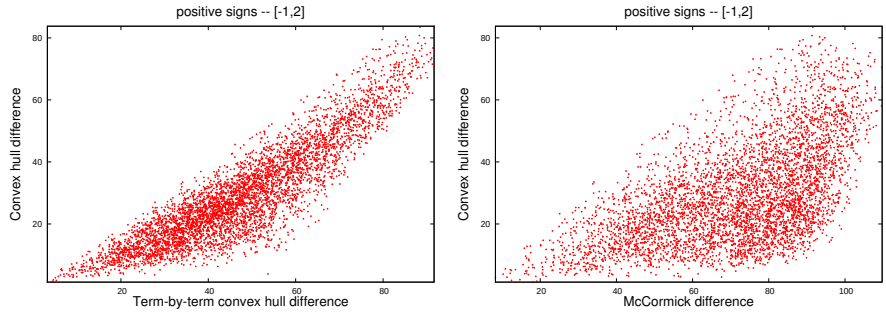

Fig. 3 Scatter plots comparing the term-by-term (left) and recursive McCormick (right) relaxation gaps to the convex hull relaxation gap for the function ϕ defined over $[-1,2]^5$.

shows scatter plots comparing the term-by-term relaxation gap to the convex hull relaxation gap (on the left) and the McCormick relaxation gap to the convex hull relaxation gap (on the right) for 5000 randomly generated points in $[1,2]^5$. Figure 3 shows the same plots for the domain $[-1,2]^5$. In both cases, the term-by-term relaxation appears significantly better than the recursive McCormick relaxation, since in the latter case the distribution of the points is shifted significantly away from the ideal case of the line with slope one.

The most interesting of these plots is the term-by-term scatter plot for the case of domain $[1,2]^5$ in Figure 2. Recall that when $\ell \geq 0$, Corollary 5 applies and hence we know the term-by-term upper relaxation yields the concave envelope. However, we have no theory suggesting the overall gap should be close to the convex hull gap. Nevertheless, the term-by-term scatter plot has the same form as the scatter plots in Figure 1 for the bilinear case, in fact with an even tighter band, suggesting that such a result might hold. In contrast, as we would expect based on the examples in Section 2.3, the results for the recursive McCormick relaxation do not suggest any such bound. Furthermore, in Figure 3 with domain $[-1,2]^5$, Theorem 5 does not apply, and thus it is not surprising that the term-by-term relaxation does significantly worse than the convex hull.

To further explore the strength of the term-by-term relaxation when $\ell \geq 0$ and all coefficients are positive, we generated 200 random multiterm multilinear functions of dimension 6, and estimated the maximum ratio of term-by-term gap to convex hull gap for each of these. We estimated this ratio by calculating the ratio at 50000 random points in the do-

main $[0,1]^6$ and taking the maximum of these. The largest estimate of the maximum ratio we found was about 1.21. This experiment, along with images like Figure 2, leads us to the following conjecture.

Conjecture 1 For multilinear functions with positive coefficients defined over $[\ell, u]$ with $\ell \geq 0$, the ratio between the term-by-term gap and the convex hull gap is uniformly bounded above by a constant.

5 Concluding remarks

We have studied the relationship between the convex hull relaxation of a multilinear function and the McCormick relaxation, obtained by relaxing individual bilinear terms. For a single product term of possibly more than two variables, we found a new condition when these relaxations are equivalent, but found that in general the McCormick relaxation can be significantly larger than the convex hull relaxation. For bilinear functions, we demonstrated that the gap between the upper and lower bounding functions from the McCormick relaxation is always within a constant factor of the gap between the concave and convex envelopes. Moreover, the maximum relative difference decreases as the coloring number of the associated graph decreases. These results, along with a result showing that the difference in these gaps is always smaller for sparser graphs, suggest that the extra benefit from using a relaxation stronger than the McCormick relaxation is likely to be most beneficial when the associated graph is dense.

This work leaves some additional theoretical and computational questions open. On the theoretical side, we believe that the approximation ratio we have provided for general bilinear functions (having both positive and negative coefficients on the terms) is not as tight as possible. We have also conjectured that using the convex hull of every term in a multilinear function having positive coefficients on all terms will yield an approximation with a gap that is within a constant factor of the gap between the concave and convex envelopes. This would be a generalization of our result for bilinear functions. On the computational side, it would be interesting to build on the ideas of [2] and use the insights gained from this paper to devise a relaxation approach for multilinear functions that yields some of the potential improvement in relaxation quality that the convex hull formulation yields while keeping the relaxation size manageable.

Acknowledgements The authors thank an anonymous referee and the special issue editors for helpful comments, and especially Pierre Bonami for pointing out an error in an earlier version of the manuscript.

References

1. Al-Khayyal, F., Falk, J.: Jointly constrained biconvex programming. Math. Oper. Res. **8**(2), 273–286 (1983)
2. Bao, X., Sahinidis, N.V., Tawarmalani, M.: Multiterm polyhedral relaxations for nonconvex, quadratically constrained quadratic programs. Optim. Methods Softw. **24**(4-5), 485–504 (2009)
3. Belotti, P., Lee, J., Liberti, L., Margot, F., W achter, A.: Branching and bounds tightening techniques for non-convex MINLP. Optim. Methods Softw. **24**, 597–634 (2009)
4. Cho, J., Raje, S., Sarrafzadeh, M.: Fast approximation algorithms on maxcut, k-coloring, and k-color ordering for VLSI applications. IEEE Trans. on Comput. **47**(11), 1253–1266 (1998)
5. Coppersmith, D., Günlük, O., Lee, J., Leung, J.: A polytope for a product of real linear functions in 0/1 variables. Tech. rep., IBM Research Report RC21568 (1999)

6. Crama, Y.: Concave extensions for nonlinear 0-1 maximization problems. Math. Program. **61**, 53–60 (1993)
7. Falk, J., Hoffman, K.: A successive underestimation method for concave minimization problems. Math. Oper. Res. **1**, 251–259 (1976)
8. Floudas, C.: Deterministic Global Optimization: Theory, Algorithms and Applications. Kluwer Academic Publishers (2000)
9. Günlük, O., Lee, J., Leung, J.: A polytope for a product of real linear functions in 0/1 variables. In: Lee, J., Leyffer, S. (eds.) Mixed Integer Nonlinear Programming, *The IMA Volumes in Mathematics and its Applications*, vol. 154, pp. 513–532. Springer (2011)
10. Haglin, D., Venkatesan, S.: Approximation and intractability results for the maximum cut problem and its variants. IEEE Trans. on Comp. **40**(1), 110–113 (1991)
11. Hochbaum, D., Megiddo, N., Naor, J., Tamir, A.: Tight bounds and 2-Approximation algorithms for integer programs with two variables per inequality. Math. Program. pp. 69–83 (1993)
12. Kahruman, S., Kolotoglu, E., Butenko, S., Hicks, I.: On greedy construction heuristics for the MAX-CUT problem. Int. J. of Comput. Sci. and Eng. **3**(3), 211–218 (2007)
13. Kajitani, Y., Cho, J., Sarrafzadeh, M.: New approximation results on graph matching and related problems. In: Mayr, E., Schmidt, G., Tinhofer, G. (eds.) Graph-Theroetic Concepts in Computer Science, *Lecture Notes in Computer Science 903*, vol. 45, pp. 343–358. Herrsching, Germany (1995)
14. McCormick, G.P.: Computability of global solutions to factorable nonconvex programs: Part I—Convex underestimating problems. Math. Program. **10**, 147–175 (1976)
15. Meyer, C., Floudas, C.: Convex envelopes for edge-concave functions. Math. Program., Ser. B **103**, 207–224 (2005)
16. Motwani, R., Raghaven, P.: Randomized Algorithms. Cambridge University Press, Cambridge, UK (1995)
17. Namazifar, M.: Strong relaxations and computations for multilinear programming. Ph.D. thesis, University of Wisconsin-Madison (2011)
18. Padberg, M.: The boolean quadric polytope: some characteristics, facets and relatives. Math. Program. **45**, 139–172 (1989)
19. Rikun, A.D.: A convex envelope formula for multilinear functions. J. Global Opt. **10**, 425–437 (1997)
20. Ryoo, H.S., Sahinidis, N.V.: Analysis of bounds for multilinear functions. J. Global Opt. **19**, 403–424 (2001)
21. Sahinidis, N.: BARON: A general purpose global optimization software package. J. Global Opt. **8**, 201–205 (1996)
22. Sherali, H.: Convex envelopes of multilinear functions over a unit hypercube and over special discrete sets. Acta Math. Vietnam. **22**, 245–270 (1997)
23. Smith, E., Pantelides, C.: A symbolic reformulation/spatial branch-and-bound algorithm for the global optimisation of nonconvex minlps. Comput. & Chem. Eng. **23**, 457–478 (1999)
24. Tawaramalani, M., Sahinidis, N.: A polyhedral branch-and-cut approach to global optimization. Math. Program. **103**, 225–249 (2005)