

# Some Results on the Strength of Relaxations of Multilinear Functions

James Luedtke · Mahdi Namazifar · Jeff Linderoth

August 13, 2010

**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, these relaxations are equivalent 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 overestimator and convex underestimator 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, provide insight into how to construct strong relaxations of multilinear functions.

**Keywords** Global optimization · Bilinear function · Multilinear function

## 1 Introduction

The construction of convex underestimators and concave overestimators for nonconvex functions plays a critical role in algorithms for globally solving nonconvex optimization problems. In this work, we focus on multilinear functions  $\phi : [\ell, u] \rightarrow \mathbb{R}$ , where

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

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 set

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

---

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

An important special case is when  $\phi$  is a *bilinear* function having  $|J_t| \leq 2$  for all  $t \in T$ .

When  $\phi(x)$  consists of a single bilinear term, McCormick [13] 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:

$$z \geq u_2 x_1 + u_1 x_2 - u_1 u_2, \quad z \geq \ell_2 x_1 + \ell_1 x_2 - \ell_1 \ell_2, \quad (2a)$$

$$z \leq u_2 x_1 + \ell_1 x_2 - \ell_1 u_2, \quad z \leq \ell_2 x_1 + u_1 x_2 - u_1 \ell_2. \quad (2b)$$

Al-Khayyal and Falk [1] showed that the convex hull of  $B$  is described by the McCormick inequalities. 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 the software BARON [19,21] and Couenne [3].

The strongest possible relaxation of  $X$ , its convex hull  $\text{conv}(X)$ , has been shown to be a polyhedron with the following characterization [6–8, 17, 20]:

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

where  $x^1, x^2, \dots, x^{2^n}$  are the vertices of  $[\ell, u]$ , and  $\Delta_{2^n}$  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.

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, in this work 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  $\phi$  consists of single product term ( $|T| = 1$ ). We first review a result of Ryoo and Sahinidis [18] 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  $\phi$  can have multiple terms. We begin by reviewing an existing result of Meyer and Floudas [14] which states that the *concave* envelope of  $\phi$  over  $x \in [0, 1]^n$  can be obtained as the sum of concave envelopes of the individual terms of  $\phi$  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  $\phi$  is bilinear. While these results are interesting, they do not say anything about how the *convex* envelope of  $\phi$  compares to the convex underestimator obtained from the McCormick relaxation. In §3.2 we focus on bilinear functions of the

form  $b(x) = \sum_{(i,j) \in E} a_{ij} x_i x_j$ , and prove that if  $a_{ij} > 0$  for all  $(i, j) \in E$ , then the gap between the McCormick concave overestimator and convex underestimator of  $b$  is within a constant factor of the gap between the concave and convex envelope of  $b$ . As a special case of our result, we obtain an alternate proof of the result of Coppersmith et. al. [5] that the McCormick relaxation is equivalent to the convex hull when graph  $G = (N, E)$  is bipartite, where  $N = \{1, \dots, n\}$ . More generally, our bound on the ratio decreases with the coloring number of  $G$ . We also consider the general bilinear case in which the coefficients are not all positive. We again obtain a bound on the relative gap between the McCormick and convex hull relaxations, but in this case the bound can be as bad as  $O(n)$ , although it is a constant if  $G$  is bipartite. We also show that as terms are removed from the bilinear function, 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.

*Notation:* Given a function  $f : D \rightarrow \mathbb{R}$ , the concave envelope of  $f$  over  $D$ , written  $\text{cav}_D[f]$ , is defined to be the minimum concave overestimator of  $f$  on  $D$ . That is,  $\text{cav}_D[f](x) \geq f(x)$  for all  $x \in D$ , and if  $g : D \rightarrow \mathbb{R}$  is any other concave function with  $g \geq f$  on  $D$ , then  $g \leq \text{cav}_D[f]$ . Similarly, the convex envelope of  $f$  over  $D$ , written  $\text{vex}_D[f]$ , is defined to be the maximum convex underestimator 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  $i$ th 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  $\text{Diag}(u)$  to be the  $n \times n$  diagonal matrix with  $\text{Diag}(u)_{ii} = u_i$ .

## 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 the 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  $\text{conv}(X)$ .

### 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 [18]. First, for  $(x_1, x_2) \in [l_1, u_1] \times [l_2, u_2]$  define  $MC_{[l_1, u_1] \times [l_2, u_2]}(x_1, x_2)$  to be the closed interval

$$MC_{[l_1, u_1] \times [l_2, u_2]}(x_1, x_2) \stackrel{\text{def}}{=} [\max\{u_2 x_1 + u_1 x_2 - u_1 u_2, \ell_2 x_1 + \ell_1 x_2 - \ell_1 \ell_2\}, \\ \min\{u_2 x_1 + \ell_1 x_2 - \ell_1 u_2, \ell_2 x_1 + u_1 x_2 - u_1 \ell_2\}]$$

and observe that  $(y_1, x_1, x_2)$  satisfies the McCormick inequalities (2) if and only if  $y_1 \in MC_{[l_1, u_1] \times [l_2, u_2]}(x_1, x_2)$ .

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  $y_2, \dots, y_n$  that satisfy  $y_i = x_i y_{i+1}$  for  $i = 1, \dots, n-1$  and  $y_n = x_n$

and then relaxing these constraints with the McCormick inequalities. This leads to a “recursive” McCormick relaxation of  $X_{[\ell,u]}$  which is the polytope define by:

$$\text{RMC}(X_{[\ell,u]}) = \left\{ (x, y) \in [\ell, u] \times \mathbb{R}^n \mid y_n = x_n, \right. \\ \left. y_i \in MC_{[\ell_i, u_i] \times [\tilde{\ell}_{i+1}, \tilde{u}_{i+1}]}(x_i, y_{i+1}), i = 1, \dots, n-1 \right\}$$

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  $i = n-1, \dots, 1$ . The variable  $y_n$  could be eliminated from the description of  $\text{RMC}(X_{[\ell,u]})$ , but we include it for notational convenience.

We are now ready to state the result of Ryoo and Sahinidis [18].

**Theorem 1 ([18])** *Let  $f(x) = \prod_{i=1}^n x_i$ . The recursive McCormick relaxation describes the convex hull of  $f$  over  $H$ , i.e.,  $\text{Proj}_{(x,y_1)}(\text{RMC}(X_H)) = \text{conv}(X_H)$ .*

As observed in [18], 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  $[\mathbf{0}, u]$ . For completeness, we provide a proof of this statement in the appendix.

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

## 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  $i = 1, \dots, n$  are  $[\tilde{\ell}_i, \tilde{u}_i] = [-1, 1]$ . Consequently, the conditions  $y_i \in MC_{[-1,1]^2}(x_i, y_{i+1})$  in the definition of  $\text{RMC}(X_{[-\mathbf{1},\mathbf{1}]})$  have the form

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

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

We will use the following simple lemma. The proof is in the appendix.

**Lemma 1** *Let  $x_1, x_2 \in [-1, 1]$ . If either  $x_1 \in \{-1, 1\}$  or  $x_2 \in \{-1, 1\}$  then*

$$\max\{-x_1 - x_2 - 1, x_1 + x_2 - 1\} = \min\{x_1 - x_2 + 1, -x_1 + x_2 + 1\}.$$

*Thus, if  $y \in MC_{[-1,1]^2}(x_1, x_2)$  then  $y = x_1x_2$ . Conversely, if  $x_1 \in (-1, 1)$  and  $x_2 \in (-1, 1)$  then*

$$\max\{-x_1 - x_2 - 1, x_1 + x_2 - 1\} < \min\{x_1 - x_2 + 1, -x_1 + x_2 + 1\}$$

*so that  $MC_{[-1,1]^2}(x_1, x_2)$  is a positive length interval.*

**Theorem 2** *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.,  $\text{Proj}_{(x,y_1)}(\text{RMC}(X_{[-\mathbf{1},\mathbf{1}]})) = \text{conv}(X_{[-\mathbf{1},\mathbf{1}]})$ .*

*Proof* We only need to prove  $\text{Proj}_{(x,y_1)}(\text{RMC}(X_{[-\mathbf{1},\mathbf{1}]})) \subseteq \text{conv}(X_{[-\mathbf{1},\mathbf{1}]})$ . We will prove that if  $(x, y)$  is an extreme point of  $\text{RMC}(X_{[-\mathbf{1},\mathbf{1}]})$  then  $y_1 = \prod_{i=1}^n x_i$ , and hence  $(x, y_1) \in X_{[-\mathbf{1},\mathbf{1}]}$ . This is sufficient to prove the

result, since this shows that every point in  $\text{Proj}(\text{RMC}(X_{[-1,1]}))$  can be written as a convex combination of points in  $X_{[-1,1]}$ .

Therefore, let  $(x, y)$  be an extreme point of  $\text{RMC}(X_{[-1,1]})$ . We consider three cases. First, if  $x \in \{-1, 1\}^n$  then Lemma 1 immediately implies  $y_i = x_i y_{i+1} = \prod_{j=i}^n x_j$  for  $i = n-1, \dots, 1$ . Second, if for some  $k$ ,  $x_i \in \{-1, 1\}$  for all  $i \neq k$ , then  $y_{k+1} = \prod_{j=k+1}^n x_j \in \{-1, 1\}$ , and hence  $y_k = x_k y_{k+1}$  by Lemma 1. Thus,  $y_k = \prod_{j=k}^n x_j$ , and because  $x_i \in \{-1, 1\}$  for all  $i < k$ , Lemma 1 implies  $y_i = \prod_{j=i}^n x_j$  for  $j = k-1, \dots, 1$ . We will show that the last case,  $|\{i \mid x_i \in (-1, 1)\}| \geq 2$ , cannot happen, since if it did  $(x, y)$  would not be an extreme point. Assume to the contrary that  $|\{i \mid x_i \in (-1, 1)\}| \geq 2$  and let  $t = \max\{j \mid x_j \in (-1, 1)\}$  and  $s = \max\{j < t \mid x_j \in (-1, 1)\}$ . Since  $x_i \in \{-1, 1\}$  for  $i = s+1, \dots, t-1$  and for  $i > t$ , an argument identical to the second case shows that  $y_i = \prod_{j=i}^n x_j$  for  $i = s+1, \dots, n$  and hence  $y_i \in \{-1, 1\}$  for  $i = t+1, \dots, n$  and  $y_i \in (-1, 1)$  for  $i = s+1, \dots, t$ . Next, because  $y_{s+1} \in (-1, 1)$  and  $x_s \in (-1, 1)$  Lemma 1 implies that either  $y_s > \max\{-x_s - y_{s+1} - 1, x_s + y_{s+1} - 1\}$  or  $y_s < \min\{x_s - y_{s+1} + 1, -x_s + y_{s+1} + 1\}$ .

Suppose first  $y_s > \max\{-x_s - y_{s+1} - 1, x_s + y_{s+1} - 1\}$ . Consider the vector  $(\Delta x, \Delta y)$  defined by

$$\Delta y = \sum_{i=s+1}^t (y_{t+1} \prod_{j=i}^{t-1} x_j) e_i, \quad \Delta x = (y_{t+1} \prod_{j=s+1}^{t-1} x_j) e_s + e_t$$

where we use the conventions that a product of an empty set is one and  $y_{n+1} = 1$  (if  $t = n$ ). We claim that there exists  $\epsilon > 0$  such that the vectors  $(x \pm \epsilon \Delta x, y \pm \epsilon \Delta y)$  are in  $\text{RMC}(X_{[-1,1]})$  which shows  $(x, y)$  is not an extreme point. Consider the vector  $(x', y') = (x + \epsilon \Delta x, y + \epsilon \Delta y)$ ; the argument for  $(x - \epsilon \Delta x, y - \epsilon \Delta y)$  is similar. As  $x_s \in (-1, 1)$  and  $x_t \in (-1, 1)$ , we can take  $\epsilon > 0$  small enough so that  $x' \in [-1, 1]$ . Next, as  $y_i \in \text{MC}_{[-1,1]^2}(x_i, y_{i+1})$  for  $i > t$  and  $i < s$ , the same holds for  $y'_i$  since  $(y_i, x_i, y_{i+1})$  are unchanged for all  $i > t$  and  $i < s$ . Next, using  $y_t = x_t y_{t+1}$ ,  $y_{t+1} = y'_{t+1}$ , and  $x'_t = x_t + \epsilon$  we have

$$y'_t = y_t + y_{t+1} \epsilon = x_t y_{t+1} + y_{t+1} \epsilon = (x_t + \epsilon) y_{t+1} = x'_t y'_{t+1}$$

which implies  $y'_t \in \text{MC}_{[-1,1]^2}(x'_t, y'_{t+1})$ . Similarly, for  $i = t-1, \dots, s+1$ , using  $y_i = x_i y_{i+1}$  and  $x'_i = x_i$  yields

$$y'_i = y_i + (y_{t+1} \prod_{j=i}^{t-1} x_j) \epsilon = x_i (y_{i+1} + (y_{t+1} \prod_{j=i+1}^{t-1} x_j) \epsilon) = x'_i y'_{i+1}$$

and hence  $y'_i \in \text{MC}_{[-1,1]^2}(x'_i, y'_{i+1})$ . Finally, for  $i = s$ , we have assumed that  $y_s > \max\{-x_s - y_{s+1} - 1, x_s + y_{s+1} - 1\}$  and hence there exists  $\epsilon > 0$  such that  $y'_s = y_s > \max\{-x'_s - y'_{s+1} - 1, x'_s + y'_{s+1} - 1\}$ . On the other hand,

$$x'_s - y'_{s+1} = x_s + \prod_{j=i}^{t-1} (y_{t+1} \prod_{j=i} x_j) \epsilon - (y_{s+1} + \prod_{j=i}^{t-1} (y_{t+1} \prod_{j=i} x_j) \epsilon) = x_s - y_{s+1}$$

and hence  $y_s \leq \min\{x_s - y_{s+1} + 1, -x_s + y_{s+1} + 1\}$  implies that  $y'_s = y_s \leq \min\{x'_s - y'_{s+1} + 1, -x'_s + y'_{s+1} + 1\}$  and hence  $y'_s \in \text{MC}_{[-1,1]^2}(x'_s, y'_{s+1})$  completing the proof that  $(x', y') \in \text{RMC}(X_{[-1,1]})$ .

The case  $y_s < \min\{x_s - y_{s+1} + 1, -x_s + y_{s+1} + 1\}$  is nearly identical, except that in this case we define  $(\Delta x, \Delta y)$  by

$$\Delta y = \sum_{i=s+1}^t (y_{t+1} \prod_{j=i}^{t-1} x_j) e_i, \quad \Delta x = -(y_{t+1} \prod_{j=s+1}^{t-1} x_j) e_s + e_t.$$

Then, again considering  $(x', y') = (x + \epsilon \Delta x, y + \epsilon \Delta y)$ , the arguments are identical except that for checking that  $y'_s \in \text{MC}_{[-1,1]^2}(x'_s, y'_{s+1})$ . In this case, existence of  $\epsilon > 0$  such that  $y'_s < \min\{x'_s - y'_{s+1} + 1, -x'_s + y'_{s+1} + 1\}$  follows from the same strict inequality holding for  $(y_s, x_s, y_{s+1})$ . The condition  $y'_s \geq \max\{-x'_s - y'_{s+1} - 1, x'_s + y'_{s+1} - 1\}$

follows from the same holding for  $(y_s, x_s, y_{s+1})$  and the observation that

$$x'_s + y'_{s+1} = x_s - \left( y_{t+1} \prod_{j=s+1}^{t-1} x_j \right) \epsilon + y_{s+1} + \left( y_{t+1} \prod_{j=s+1}^{t-1} x_j \right) \epsilon = x_s + y_{s+1}. \square$$

Using arguments that are identical to the proof of Corollary 1, we obtain the following generalization.

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

### 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 either of these conditions is 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

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

The relation  $\text{rmcgap}_D[f](x) \geq \text{chgap}_D[f](x)$  always holds, and equality holds when either  $\ell = \mathbf{0}$  or  $\ell = -u$ . The examples we give show that  $\text{rmcgap}_D[f](x) \gg \text{chgap}_D[f](x)$  is possible, and the difference can be arbitrarily large.

First, let  $D_3 = [1, u]^3$  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  $D_3$  is  $\hat{x} = \frac{1}{2}(1, u, 1) + \frac{1}{2}(u, u, 1)$ . Thus,  $\text{vex}_{D_3}[f](\hat{x}) = \text{cav}_{D_3}[f](\hat{x})$  so  $\text{chgap}_{D_3}[f](\hat{x}) = 0$ . Next consider the recursive McCormick relaxation of  $f$  over  $D_3$ . It is possible to check that  $\text{rmcu}_{D_3}[f](\hat{x}) = u^2 + \frac{1-u}{2}$  and  $\text{rmcl}_{D_3}[f](\hat{x}) = u + \frac{u-1}{2}$ , and therefore

$$\text{rmcgap}_{D_3}[f](\hat{x}) = \text{rmcu}_{D_3}[f](\hat{x}) - \text{rmcl}_{D_3}[f](\hat{x}) = u^2 - u > 0.$$

Since  $\text{chgap}_{D_3}[f](\hat{x}) = 0$ , this example shows that, if we let  $u \rightarrow \infty$ , 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  $\hat{x} = (2, \dots, 2, 0, 0, 2)$ . Again, the only way  $\hat{x}$  can be written as a convex combination of vertices of  $D_n$  is  $\hat{x} = \frac{1}{2}(2, \dots, 2, -2, 0, 2) + \frac{1}{2}(2, \dots, 2, 2, 0, 2)$  and hence  $\text{vex}_{D_n}[f](\hat{x}) = \text{cav}_{D_n}[f](\hat{x})$  so  $\text{chgap}_{D_n}[f](\hat{x}) = 0$ . On the other hand, if we consider the recursive McCormick relaxation of  $f$  over  $D_n$ , it can be verified that for  $n \geq 3$ ,  $\text{rmcu}_{D_n}[f](\hat{x}) = 2^n$  and  $\text{rmcl}_{D_n}[f](\hat{x}) = -2^n$  and hence  $\text{rmcgap}_{D_n}[f](\hat{x}) = 2^{n+1}$ . Thus, by letting  $n \rightarrow \infty$ , we see that even if the bounds  $\ell$  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

$$\phi(x) = \sum_{t \in T} a_t \prod_{j \in J_t} x_j, \quad (4)$$

defined over  $x \in [\ell, u]$ , and study concave overestimators and convex underestimators of  $\phi$  over  $[\ell, u]$ . In Section 3.1 we will focus on concave envelopes, and study cases in which the concave envelope of  $\phi$  can be written as a sum of concave envelopes of the individual terms. These results mostly follow directly from results in [6, 14], but we state them here 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 simple McCormick overestimator and underestimator of  $\phi$  is uniformly within a constant of the gap between the concave and convex envelopes of  $\phi$ .

Recall that the concave and convex envelopes of  $\phi$  have the following representations [17, 20]:

$$\text{cav}_{[\ell, u]}[\phi](x) = \max_{\lambda} \left\{ \sum_{k=1}^{2^n} \lambda_k \phi(x^k) \mid \sum_{k=1}^{2^n} \lambda_k x^k = x, \lambda \in \Delta_{2^n} \right\} \quad (5)$$

$$\text{vex}_{[\ell, u]}[\phi](x) = \min_{\lambda} \left\{ \sum_{k=1}^{2^n} \lambda_k \phi(x^k) \mid \sum_{k=1}^{2^n} \lambda_k x^k = x, \lambda \in \Delta_{2^n} \right\} \quad (6)$$

where  $x^k, k = 1, \dots, 2^n$  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 [14].

**Theorem 3 ([14])** *Let  $\phi : H \rightarrow \mathbb{R}$  be as defined in (4), 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  $\phi$  is given by the sum of concave envelopes of  $f_t$ :*

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

This result also holds for  $x \in [\ell, u]$  as long as  $\ell \geq \mathbf{0}$ .

**Corollary 3** *Let  $\ell, u \in \mathbb{R}^n$  satisfy  $\mathbf{0} \leq \ell \leq u$  and let  $\phi : [\ell, u] \rightarrow \mathbb{R}$  be as defined in (4), 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  $\phi$  over  $[\ell, u]$  is given by the sum of concave envelopes of  $f_t$ :*

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

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

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

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 will be a product of  $\ell_j$  and  $(u_j - \ell_j)$  terms and  $\ell_j \geq 0$ . Now, let  $\phi'_t(x') = \sum_{k \in K_t} a'_k f'_k(x')$  for  $t \in T$ .

Applying Theorem 3 twice then yields

$$\text{cav}_H[\phi'](x') = \sum_{t \in T} a_t \sum_{k \in K_t} a'_k \text{cav}_H[f'_k](x') = \sum_{t \in T} a_t \text{cav}_H[\phi'_t](x') \quad \forall x' \in H. \quad (7)$$

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

$$\text{cav}_H[\phi'_t](x') = \text{cav}_{[\ell, u]}[f_t](\text{Diag}(u - \ell)x' + \ell), \quad \forall x' \in H. \quad (8)$$

Now, let  $x \in [\ell, u]$  and let  $x' = \text{Diag}(u - \ell)^{-1}(x - \ell)$  and  $y' = \text{cav}_H[\phi'](x')$ . 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  $\tilde{x}^k, k = 1, \dots, 2^n$  are the vertices of  $H$ . Then, observing that  $x^k = \text{Diag}(u - \ell)\tilde{x}^k + \ell$ , for  $k = 1, \dots, 2^n$  are the vertices of  $[\ell, u]$  we have

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

and so  $\lambda$  is feasible to the linear program (5) defining  $\text{cav}_{[\ell, u]}[\phi]$ . Also, the objective value of  $\lambda$  in (5) is

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

where the first second-to-last equality follows from (7) and the last equality follows from (8). This proves

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

and completes the proof as the reverse inequality is immediate.  $\square$

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

**Corollary 4** *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 overestimator:*

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

### 3.2 Approximation results for bilinear functions

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

$$b(x) = \sum_{(i,j) \in E} a_{ij}x_i x_j \quad (9)$$

for  $x \in H$  where  $E$  is a subset of pairs of indices in  $N = \{1, \dots, n\}$ . Specifically, the McCormick overestimator is

$$\text{mcu}_H[b](x) = \max\left\{ \sum_{(i,j) \in E} a_{ij}y_{ij} \mid (x, y) \in P \right\}$$

and the McCormick underestimator is

$$\text{mcl}_H[b](x) = \min\left\{ \sum_{(i,j) \in E} a_{ij}y_{ij} \mid (x, y) \in P \right\}$$

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  $x_i x_j$ .

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

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

$\text{mcgap}_H[b](x)$  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  $\text{chgap}_H[b](x)$ . In this section, we will show that under certain conditions,  $\text{mcgap}_H[b](x)$  is uniformly close to  $\text{chgap}_H[b](x)$  over  $x \in H$ .

We begin in Section 3.2.1 by reviewing some existing results and establishing some new results needed for proving out 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 4.

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 4 ([16])** *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 [16], Theorem 4 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 5 ([12])** *Consider any graph  $G = (V, E)$  having  $|V|$  even and weights  $w_e$  for  $e \in E$ . There exists a matching  $M \subseteq E$ , with*

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

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 [15]). 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 [9] using Theorem 5 in place of the (weaker) bound on the size of a matching used in [4]. (See also the discussion in [11]).

**Corollary 5** *Let  $G = (V, E)$  be a graph with  $|V|$  even and weights  $w_e$  for  $e \in E$ . Then there exists a cut  $C \subseteq E$  in  $G$  having*

$$w(C) \geq \frac{1}{2}w(E) + \frac{1}{2(|V| - 1)}w^+(E).$$

*Proof* By Theorem 5, there exists a matching  $M$  in the graph  $(V, E^+)$  with  $w(M) \geq w(E^+)/(|V| - 1) = w^+(E)/(|V| - 1)$ . 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  $e = (i, j) \in M$ , 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  $1/2$ . Thus, with probability 1, every edge in  $M$  is 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:

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

This implies there exists a cut that achieves the value of the expected weight of this random cut.  $\square$

This result can be strengthened further for graphs that have a small coloring number.

**Lemma 2** *Let  $G = (V, E)$  be a graph with  $\chi(G)$  even, and weights  $w_e$  for  $e \in E$ . Then there exists cuts  $C^+$  and  $C^-$  in  $G$  with*

$$\begin{aligned} w(C^+) &\geq \frac{1}{2}w(E) + \frac{1}{2(\chi(G) - 1)}w^+(E), \\ w(C^-) &\leq \frac{1}{2}w(E) + \frac{1}{2(\chi(G) - 1)}w^-(E). \end{aligned}$$

*Proof* We prove the existence of the cut  $C^+$ ; the existence of  $C^-$  can be achieved by applying the  $C^+$  result to a graph with weights  $\bar{w}_e = -w_e$ . Let  $\chi = \chi(G)$  and let  $S_1, \dots, S_\chi$  be a partition of  $V$  such that  $\gamma(S_i) = \emptyset$  for all  $i = 1, \dots, \chi$ . (I.e., these sets define a coloring of size  $\chi$ .) Define a complete graph  $G'$  with vertices  $V' = \{1, \dots, \chi\}$ , 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 5 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  $w(C) = \bar{w}(C')$ ,  $w^+(C) = \bar{w}^+(C')$  and  $C$  is a cut in  $G$ , this proves the result.  $\square$

Due to Theorem 4, vectors  $x$  that are  $\{0, 1/2, 1\}$ -valued will play an important role in our analysis. We therefore determine  $\text{mcgap}_H[b](x)$  and find bounds on  $\text{cav}_H[b](x)$  and  $\text{vex}_H[b](x)$  for such vectors.

**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  $T_f = \{i \in N \mid x_i = 1/2\}$ . Then*

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

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

$$\text{mcgap}_H[b](x) = \sum_{(i,j) \in E} |a_{ij}| (\min\{x_i, x_j\} - \max\{x_i + x_j - 1, 0\}). \quad (10)$$

Indeed,

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

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 (10).  $\square$

**Lemma 4** *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,*

$$\text{vex}_H[b](x) \leq a(\gamma(T_1)) + \frac{1}{2}a(\delta(T_1, T_f)) + \frac{1}{4}a(\gamma(T_f)) - \frac{1}{4(\chi(G) - 1)}a^+(\gamma(T_f)) \quad (11)$$

and

$$\text{cav}_H[b](x) \geq a(\gamma(T_1)) + \frac{1}{2}a(\delta(T_1, T_f)) + \frac{1}{4}a(\gamma(T_f)) - \frac{1}{4(\chi(G) - 1)}a^-(\gamma(T_f)). \quad (12)$$

*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 (6) defining  $\text{vex}_H[b](x)$  as follows:

$$\text{vex}_H[b](x) = \min_{\lambda \in \Delta_{2^n}} \sum_{S \in \mathcal{S}} a(\gamma(S)) \lambda_S \quad (13a)$$

$$\text{s.t. } \sum_{S \in \mathcal{S}_i} \lambda_S = x_i, \quad i = 1, \dots, n. \quad (13b)$$

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  $U_1 \cap U_2 = \emptyset$ ). Since the coloring number of  $G_f$  will be no larger than the coloring number of  $G$ , Lemma 2 implies

$$a(C) \geq \frac{1}{2}a(\gamma(T_f)) + \frac{1}{2(\chi(G) - 1)}a^+(\gamma(T_f)). \quad (14)$$

Now, let  $S_1 = T_1 \cup U_1$  and  $S_2 = T_1 \cup U_2$ , and construct a solution to (13) 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 (13b) is satisfied as well. Thus, because  $\lambda$  is one feasible solution to (13),

$$\text{vex}_H[b](x) \leq \frac{1}{2}(a(\gamma(S_1)) + a(\gamma(S_2))). \quad (15)$$

Next, recalling 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

$$\begin{aligned} & a(\gamma(S_1)) + a(\gamma(S_2)) \\ &= 2a(\gamma(T_1)) + a(\delta(T_1, T_f)) + a(\gamma(T_f)) - a(C) \\ &\leq 2a(\gamma(T_1)) + a(\delta(T_1, T_f)) + \frac{1}{2}a(\gamma(T_f)) - \frac{1}{2(\chi(G) - 1)}a^+(\gamma(T_f)) \end{aligned}$$

where the inequality follows from (14). Using this in (15) yields (11).

The proof of (12) is similar except that we use Lemma 2 to show there exists a cut  $C^-$  such that

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

This cut can then be used to construct a feasible solution to the maximization problem defining  $\text{cav}_H[b](x)$  with objective value equal to the lower bound in (12).  $\square$

### 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 6** *Let  $G = (N, E)$  have a coloring of size  $\chi$ , and let  $b(x)$  be a bilinear function of the form (9) with  $a_{ij} > 0$  for all  $(i, j) \in E$ . Then if  $\chi$  is even,*

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

and if  $\chi$  is odd,

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

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].

*Proof* We prove the case where  $\chi$  is even. The case where  $\chi$  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

$$\min_{x \in H} \{K \text{chgap}_H[b](x) - \text{mcgap}_H[b](x)\} \geq 0. \quad (16)$$

Next, because  $a_{ij} > 0$  for all  $(i, j) \in E$ , Theorem 3 applies and hence  $\text{cav}_H[b](x) = \text{mcu}_H[b](x)$ . Using this, the definitions of  $\text{chgap}_H[b]$  and  $\text{mcgap}_H[b]$ , and expanding the definition of  $\text{mcl}_H[b](x)$ , the minimization problem in (16) is equivalent to:

$$\min \left\{ (K - 1) \text{cav}_H[b](x) - K \text{vex}_H[b](x) + \sum_{(i,j) \in E} a_{ij} y_{ij} \mid (x, y) \in P \right\}$$

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 4. Then, because  $\text{cav}_H[b](x)$  and  $-\text{vex}_H[b](x)$  are concave functions, the above problem is a concave minimization problem over a polytope, and hence achieves its minimum at an extreme point. Theorem 4 then implies that it is sufficient to prove

$$K \text{chgap}_H[b](x) - \text{mcgap}_H[b](x) \geq 0 \quad (17)$$

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 3 then implies

$$\text{mcgap}_H[b](x) = \frac{1}{2} \sum_{(i,j) \in \gamma(T_f)} |a_{ij}| = \frac{1}{2} a(\gamma(T_f)). \quad (18)$$

Next, again using Theorem 3,

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

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 (11) from Lemma 4 and (18) yields

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

where in using (11) we have also used  $a^+(\gamma(T_f)) = a(\gamma(T_f))$  since here  $a_{ij} \geq 0$  for all  $(i, j)$ . Rearranging yields

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

and so indeed (17) holds.  $\square$

### 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 7** *Let  $G = (N, E)$  have a coloring of size  $\chi$ , and let  $b(x)$  be a bilinear function of the form (9) over  $x \in H$ . Then if  $\chi$  is even,*

$$\text{mcgap}_H[b](x) \leq 2(\chi - 1) \text{chgap}_H[b](x) \quad \forall x \in H, \quad (19)$$

and if  $\chi$  is odd,

$$\text{mcgap}_H[b](x) \leq 2\chi \text{chgap}_H[b](x) \quad \forall x \in H.$$

*Proof* As in the proof of Theorem 6, we restrict attention to the case where  $\chi$  is even. First, for  $x_i, x_j \in [0, 1]$  observe that

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

Thus, using this in (10) we can write  $\text{mcgap}_H[b](x)$  as

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

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 [10] imply that all vertices of  $Q$  are  $\{0, 1/2, 1\}$ -valued.

Now, we need to prove

$$\min_{x \in H} \{2(\chi - 1) \text{chgap}_H[b](x) - \text{mcgap}_H[b](x)\} \geq 0.$$

This minimization problem is equivalent to:

$$\min\left\{2(\chi - 1) \text{chgap}_H[b](x) - \sum_{(i,j) \in E} |a_{ij}| z_{ij} \mid (x, z) \in Q\right\}$$

Since  $\text{chgap}_H[b](x)$  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 6, it is sufficient to show that (19) holds for  $\{0, 1/2, 1\}$ -valued  $x$ .

Thus, let  $x$  be any  $\{0, 1/2, 1\}$ -valued vector. Using Lemma 4 to bound both  $\text{vex}_H[b](x)$  and  $\text{cav}_H[b](x)$  yields

$$\begin{aligned} \text{chgap}_H[b](x) &= \text{cav}_H[b](x) - \text{vex}_H[b](x) \\ &\geq \frac{1}{4(\chi - 1)} \left( a^+(\gamma(T_f)) - a^-(\gamma(T_f)) \right) \\ &= \frac{1}{4(\chi - 1)} \sum_{(i,j) \in E} |a_{ij}| = \frac{1}{2(\chi - 1)} \text{mcgap}_H[g](x) \end{aligned}$$

by Lemma 3, completing the proof.  $\square$

The bound in Theorem 7 is significantly weaker than Theorem 6 which provides a constant approximation guarantee; in this case, the approximation factor can be as bad as  $2n$ . 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. However, for bipartite graphs, in which the coloring number is 2, the result yields a constant factor 2 which is tight.

*Example 1* 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  $\text{mcgap}_H[b](x) = (1/2) \sum_{(i,j) \in E} |a_{ij}| = 2$ . For  $\text{cav}_H[b](x)$ , the optimal value sets  $\lambda_{\{1,3\}} = \lambda_{\{2,4\}} = 1/2$  and achieves value  $(1/2)(a_{13} + a_{24}) = 1$  and for  $\text{vex}_H[b](x)$  the optimal value sets  $\lambda_{\{1,4\}} = \lambda_{\{2,3\}} = 1/2$  and achieves the value  $(1/2)(a_{14} + a_{23}) = 0$ . Thus,  $2 \text{chgap}_H[b](x) = 2 = \text{mcgap}_H[b](x)$ .

Theorems 6 and 7 both provide a worst-case approximation guarantee 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. Given a graph  $G = (V, E)$

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  $E'$ :

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

**Theorem 8** *Let  $E' \subseteq E$ . Then, for any  $x \in H$ ,*

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

*Proof* We will prove the equivalent inequality:

$$\text{mgap}_H[b_E](x) - \text{mgap}_H[b_{E'}](x) \geq \text{chgap}_H[b_E](x) - \text{chgap}_H[b_{E'}](x). \quad (20)$$

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 an inductive argument.

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

$$\text{mgap}_H[b_E](x) - \text{mgap}_H[b_{E'}](x) = |a_{kl}| (\max\{x_k + x_l - 1, 0\} - \min\{x_k, x_l\}). \quad (21)$$

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

$$\text{cav}_H[b_E](x) = \max_{\lambda \in \Delta_{2^n}} \sum_{S \in \mathcal{S}} a(\gamma^E(S)) \lambda_S \quad (22a)$$

$$\text{s.t.} \quad \sum_{S \in \mathcal{S}_i} \lambda_S = x_i, \quad i = 1, \dots, n \quad (22b)$$

where we have made dependence on the edge set  $E$  explicit:  $\gamma^E(S) = \{(i, j) \in E \mid i \in S, j \in S\}$ . Let  $\lambda^E$  be an optimal solution to (22). Clearly,  $\lambda^E$  is also a feasible solution to the problem (22) when  $E'$  replaces  $E$ . Thus,

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

But, (22b) implies  $\sum_{S \in \mathcal{S}_k \cap \mathcal{S}_l} \lambda_S^E \leq x_k$  and  $\sum_{S \in \mathcal{S}_k \cap \mathcal{S}_l} \lambda_S^E \leq x_l$  and hence,

$$\text{cav}_H[b_E](x) - \text{cav}_H[b_{E'}](x) \leq a_{kl} \min\{x_k, x_l\}. \quad (23)$$

Now let  $\lambda^E$  be an optimal solution to the linear program defining  $\text{vex}_H[b_E](x)$ , which is (22) with  $\max$  replaced by  $\min$ . As  $\lambda^E$  is also feasible to the LP defining  $\text{vex}_H[b_{E'}](x)$ , we have, similar to the argument for

$\text{cav}_H$ ,

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

Next, (22b) implies

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

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

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

Combining this with (23) implies

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

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  $\text{vex}_H[b_E](x) - \text{vex}_H[b_{E'}](x)$  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  $\text{cav}_H[b_E](x) - \text{cav}_H[b_{E'}](x)$ .  $\square$

## 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 6) and the mixed sign coefficients case (Theorem 7). Also, inspired by Theorem 8, 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:  $\max_{x \in H} \{\text{mcgap}_H[b](x) / \text{chgap}_H[b](x)\}$ . This maximum was found by calculating  $\text{mcgap}_H[b](x)$  and  $\text{chgap}_H[b](x)$  for all  $3^7$   $\{0, 1/2, 1\}$ -valued points in  $H$ .

Table 1 displays the results summarized by coloring number. For each coloring number from 2 – 7, 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 6 is tight for coloring number up to 7. Also, the vast majority of the randomly

$\chi$	Positive Coefficients			Mixed-Sign Coefficients		
	Max Ratio			Max Ratio		
	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.

generated graphs achieved this worst-case bound. In contrast, when the coefficients have mixed-sign, the bound of Theorem 7 does not appear tight, except for the case of coloring number 2, which we have already seen is tight in Example 1. In addition, even for  $\chi = 2$ , although we did generate a graph that achieved the bound of 2, the majority of graphs still had worst-case ratio of 1. As  $\chi$  increases, the worst-case ratio, while exceeding 2, does not appear to grow linearly with  $\chi$  as suggested in Theorem 7, suggesting that a constant-factor approximation may also be possible for bilinear functions having mixed-sign coefficients.

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 8 that the McCormick relaxation becomes relatively worse compared to the convex hull relaxation for denser graphs.

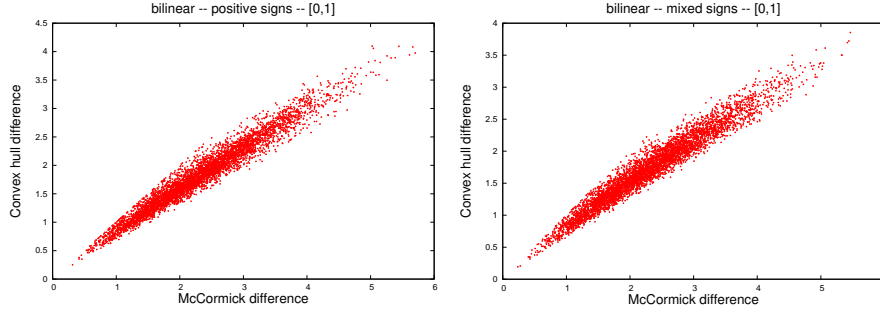
density	Positive Coefficients			Mixed-Sign Coefficients		
	Max Ratio			Max Ratio		
	avg	max	mode(%)	avg	max	mode(%)
0.0–0.1	0.000	0.000	0.000(100)	0.000	0.000	0.000 (100)
0.1–0.2	1.000	1.000	1.000(100)	1.000	1.000	1.000 (100)
0.2–0.3	1.049	1.500	1.000(90.1)	1.090	2.000	1.000 (83.5)
0.3–0.4	1.365	1.500	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.

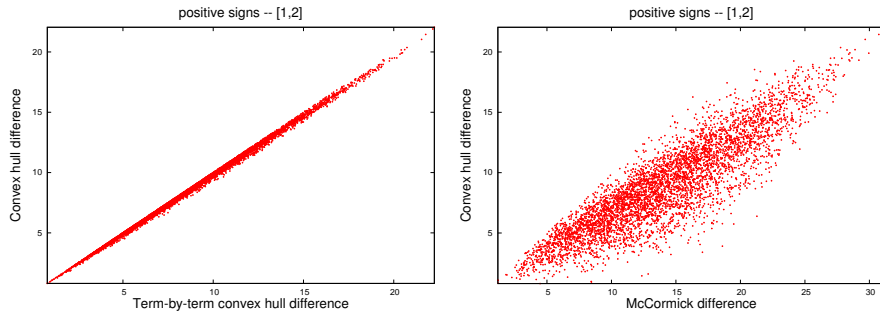
Next, we illustrate the results of Theorems 6 and 7 graphically. Consider the following two bilinear functions defined on  $x \in [0, 1]^7$ , the first having all positive coefficients and the second having mixed-signs:

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

For each of these functions, we randomly generated 5000 points uniformly in  $[0, 1]^7$  and for each point  $x^k$  calculated  $\text{mcgap}_H[b](x^k)$  and  $\text{chgap}_H[b](x^k)$ . We then construct a scatter plot of the points  $(\text{mcgap}_H[b](x^k), \text{chgap}_H[b](x^k))$ ,  $k = 1, \dots, 5000$ . These plots are shown in Figure 1. Since  $\text{mcgap}_H[b](x) \geq \text{chgap}_H[b](x)$  always holds, all of these points lie below the line of slope one originating at the origin. Moreover,



**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).



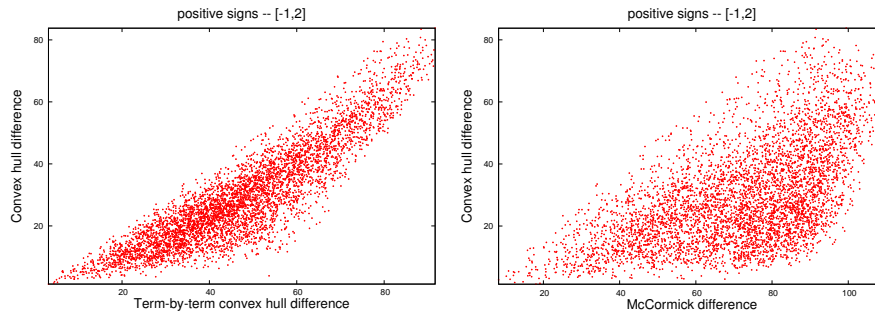
**Fig. 2** Scatter plots comparing the term-by-term (left) and recursive McCormick (right) relaxation gaps to the convex hull relaxation gap for the function  $\phi$  defined over  $[1, 2]^5$ .

in both cases, all points generated lie above a line having smaller slope, illustrating the worst-case approximation ratio. In fact, it was by looking at these plots that we first conjectured that such an approximation result would hold.

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. Corollary 3 states that if  $\ell \geq 0$  and the coefficients on all terms are positive, the concave overestimator 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, 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.



**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  $\phi$  defined over  $[-1, 2]^5$ .

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 3 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 3 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 domain  $[0, 1]^6$  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, which we have not been able to prove.

*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 estimators 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 is possible when the coloring number of the associated

graph is larger than two. 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.

## 6 Appendix

### 6.1 Proof of Corollary 1

We only need to prove  $\text{Proj}_{(x,y_1)}(\text{RMC}(X_{[\mathbf{0},u]})) \subseteq \text{conv}(X_{[\mathbf{0},u]})$ . Let  $(x, y_1) \in \text{Proj}_{(x,y_1)}(\text{RMC}(X_{[\mathbf{0},u]}))$ ,  $\bar{y}_1 = (\prod_{i=1}^n u_i)^{-1} y_1$ , and  $D_u = \text{Diag}(u)$ . We claim that  $(D_u^{-1}x, \bar{y}_1) \in \text{Proj}_{(x,y_1)}(\text{RMC}(X_H))$ . Clearly,  $D_u^{-1}x \in H$ . Let  $y_2, \dots, y_n$  be such that  $(x, y) \in \text{RMC}(X_{[\mathbf{0},u]})$  and let  $\bar{y}_i = y_i (\prod_{j=i}^n u_j)^{-1}$ ,  $i = 1, \dots, n$ . Then, it is easy to check that  $(D_u^{-1}x, \bar{y}) \in \text{RMC}(X_H)$ . Then, because  $(D_u^{-1}x, \bar{y}_1) \in \text{Proj}_{(x,y_1)}(\text{RMC}(X_H))$  Theorem 1 implies there exists  $\lambda \in \Delta_{2^n}$  such that  $\sum_k \lambda_k (x^k, y^k) = (D_u^{-1}x, \bar{y}_1)$  where  $x^k, k = 1, \dots, 2^n$  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  $X_{[\mathbf{0},u]}$ .  $\square$

### 6.2 Proof of Lemma 1

First suppose  $x_1 = 1$ . Then, because  $x_2 \in [-1, 1]$ ,  $\max\{-x_1 - x_2 - 1, x_1 + x_2 - 1\} = x_1 + x_2 - 1 = x_2$  and  $\min\{x_1 - x_2 + 1, -x_1 + x_2 + 1\} = -x_1 + x_2 + 1 = x_2$ . Similarly, if  $x_1 = -1$ ,  $\max\{-x_1 - x_2 - 1, x_1 + x_2 - 1\} = -x_1 - x_2 - 1 = -x_2$  and  $\min\{x_1 - x_2 + 1, -x_1 + x_2 + 1\} = x_1 - x_2 + 1 = -x_2$ . An identical argument works if  $x_2 \in \{-1, 1\}$  proving the first part of the claim.

Now suppose  $x_1 \in (-1, 1)$  and  $x_2 \in (-1, 1)$ . First suppose  $x_1 + x_2 > 0$ . Then,  $\max\{-x_1 - x_2 - 1, x_1 + x_2 - 1\} = x_1 + x_2 - 1 < x_1 < x_1 - x_2 + 1$  and also  $\max\{-x_1 - x_2 - 1, x_1 + x_2 - 1\} = x_1 + x_2 - 1 < x_2 < -x_1 + x_2 + 1$  and hence  $\max\{-x_1 - x_2 - 1, x_1 + x_2 - 1\} < \min\{x_1 - x_2 + 1, -x_1 + x_2 + 1\}$ . If  $x_1 + x_2 \leq 0$ , then  $\max\{-x_1 - x_2 - 1, x_1 + x_2 - 1\} = -x_1 - x_2 - 1 < -x_2 < x_1 - x_2 + 1$  and also  $\max\{-x_1 - x_2 - 1, x_1 + x_2 - 1\} = -x_1 - x_2 - 1 < -x_1 < -x_1 + x_2 + 1$  and hence again  $\max\{-x_1 - x_2 - 1, x_1 + x_2 - 1\} < \min\{x_1 - x_2 + 1, -x_1 + x_2 + 1\}$ .  $\square$

### 6.3 Proof of Corollary 4

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

$$\begin{aligned} b'(x') &= \phi(\text{Diag}(u - \ell)x' + \ell) = \sum_{(i,j) \in E} a_{ij} ((u_i - \ell_i)x'_i + \ell_i) ((u_j - \ell_j)x'_j + \ell_j) \\ &= f'(x') + L(x'), \end{aligned}$$

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

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

where the second equality follows Theorem 3 and the last equality follows because for  $f(x_1, x_2) = x_1 x_2$ ,  $\text{cav}_{[0,1]^2}[f](x_1, x_2) = \min\{x_1, x_2\}$ . It is not hard to show that also

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

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

$$\begin{aligned} \text{cav}_{[\ell, u]}[b](x) &= \text{cav}_H[b'](x') \\ &= \sum_{(i, j) \in E} a_{ij} (u_i - \ell_i) (u_j - \ell_j) \min\{x'_i, x'_j\} + L(x') \\ &= \sum_{(i, j) \in E} a_{ij} \min\{u_j x_i + \ell_i x_j - \ell_i u_j, \ell_j x_i + u_i x_j - u_i \ell_j\} \end{aligned}$$

where the last equation follows because for each  $(i, j) \in E$ ,

$$\begin{aligned} &(u_i - \ell_i) (u_j - \ell_j) \min\{x'_i, x'_j\} + \ell_j (u_i - \ell_i) x'_i + \ell_i (u_j - \ell_j) x'_j + \ell_i \ell_j \\ &= \min\{(u_j - \ell_j)(x_i - \ell_i), (u_i - \ell_i)(x_j - \ell_j)\} + \ell_j (x_i - \ell_i) + \ell_i (x_j - \ell_j) + \ell_i \ell_j \\ &= \min\{u_j x_i + \ell_i x_j - \ell_i u_j, \ell_j x_i + u_i x_j - u_i \ell_j\}. \end{aligned}$$

□

## 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. *Optimization Methods and Software* **24**(4-5), 485–504 (2009)
3. Belotti, P., Lee, J., Liberti, L., Margot, F., Wächter, 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 Transactions on Computers* **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. *Mathematical Programming* **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. Haglin, D., Venkatesan, S.: Approximation and intractability results for the maximum cut problem and its variants. *IEEE Transactions on Computers* **40**(1), 110–113 (1991)
10. 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)
11. Kahruman, S., Kolotoglu, E., Butenko, S., Hicks, I.: On greedy construction heuristics for the MAX-CUT problem. *International Journal of Computational Science and Engineering* **3**(3), 211–218 (2007)
12. Kajitani, Y., Cho, J., Sarrafzadeh, M.: New approximation results on graph matching and related problems. In: Mayr, E., Schmidt, G., Tinhofer, G. (eds.) *Graph-Theoretic Concepts in Computer Science, Lecture Notes in Computer Science 903*, vol. 45, pp. 343–358. Herrsching, Germany (1995)
13. McCormick, G.P.: Computability of global solutions to factorable nonconvex programs: Part I—Convex underestimating problems. *Mathematical Programming* **10**, 147–175 (1976)
14. Meyer, C., Floudas, C.: Convex envelopes for edge-concave functions. *Math. Program., Ser. B* **103**, 207–224 (2005)
15. Motwani, R., Raghavan, P.: *Randomized Algorithms*. Cambridge University Press, Cambridge, UK (1995)
16. Padberg, M.: The boolean quadric polytope: some characteristics, facets and relatives. *Math. Program.* **45**, 139–172 (1989)
17. Rikun, A.D.: A convex envelope formula for multilinear functions. *Journal of Global Optimization* **10**, 425–437 (1997)
18. Ryoo, H.S., Sahinidis, N.V.: Analysis of bounds for multilinear functions. *Journal of Global Optimization* **19**, 403–424 (2001)
19. Sahinidis, N.: BARON: A general purpose global optimization software package. *J. Global Opt.* **8**, 201–205 (1996)
20. Sherali, H.: Convex envelopes of multilinear functions over a unit hypercube and over special discrete sets. *Acta Mathematica Vietnamica* **22**, 245–270 (1997)
21. Tawaramalani, M., Sahinidis, N.: A polyhedral branch-and-cut approach to global optimization. *Math. Program.* **103**, 225–249 (2005)