Quadratic reformulations of nonlinear binary optimization problems

Martin Anthony[*], Endre Boros[†], Yves Crama[‡], Aritanan Gruber[§]

Abstract

Very large nonlinear unconstrained binary optimization problems arise in a broad array of applications. Several exact or heuristic techniques have proved quite successful for solving many of these problems when the objective function is a quadratic polynomial. However, no similarly efficient methods are available for the higher degree case. Since high degree objectives are becoming increasingly important in certain application areas, such as computer vision, various techniques have been recently developed to reduce the general case to the quadratic one, at the cost of increasing the number of variables by introducing additional auxiliary variables. In this paper we initiate a systematic study of these *quadratization* approaches. We provide tight lower and upper bounds on the number of auxiliary variables needed in the worst-case for general objective functions, for bounded-degree functions, and for a restricted class of quadratizations. Our upper bounds are constructive, thus yielding new quadratization procedures. Finally, we completely characterize all "minimal" quadratizations of negative monomials.

[*] Department of Mathematics, London School of Economics, UK. m.anthony@lse.ac.uk

[†] MSIS Department and RUTCOR, Rutgers University, NJ, USA. endre.boros@rutgers.edu

[‡] QuantOM, HEC Management School, University of Liège, Belgium. yves.crama@ulg.ac.be

[§] MSIS Department and RUTCOR, Rutgers University, NJ, USA. aritanan.gruber@gmail.com

1 Introduction

A *pseudo-Boolean function* is a real-valued function $f(x) = f(x_1, x_2, \ldots, x_n)$ of n binary variables, that is, a mapping from $\{0,1\}^n$ to $\mathbb{R}$. It is well-known that every pseudo-Boolean function can be uniquely represented as a multilinear polynomial in its variables. *Nonlinear binary optimization problems*, or *pseudo-Boolean optimization* (PBO) *problems*, of the form

$$\min\{f(x) : x \in \{0,1\}^n\}, \tag{1}$$

where $f(x)$ is a pseudo-Boolean function represented by its multilinear expression, have attracted the attention of numerous researchers for more than 50 years. These problems are notoriously difficult, as they naturally encompass a broad variety of models such as maximum satisfiability, max cut, graph coloring, simple plant location, problems in computer vision, and so on.

In recent years, several authors have revisited an approach to the solution of PBO initially proposed by Rosenberg [42]. This approach reduces PBO to its quadratic case (QPBO) by relying on the following concept.

Definition 1 *For a pseudo-Boolean function $f(x)$ on $\{0,1\}^n$, we say that $g(x,y)$ is a* quadratization *of f if $g(x,y)$ is a quadratic polynomial depending on x and on m* auxiliary *binary variables $y_1, y_2, \ldots, y_m$, such that*

$$f(x) = \min\{g(x,y) : y \in \{0,1\}^m\} \quad \textit{for all } x \in \{0,1\}^n. \tag{2}$$

It is quite obvious that, if $g(x,y)$ is a quadratization of f, then

$$\min\{f(x) : x \in \{0,1\}^n\} = \min\{g(x,y) : x \in \{0,1\}^n, y \in \{0,1\}^m\},$$

so that the minimization of f is reduced through this transformation to the QPBO problem of minimizing $g(x,y)$. Rosenberg [42] has observed that every pseudo-Boolean function $f(x)$ has a quadratization, and that a quadratization can be efficiently computed from the multilinear polynomial expression of f. Of course, quadratic PBO problems remain NP-hard, but this special class of problems has been thoroughly scrutinized for the past 30 years or so, and much progress has been made in solving large instances of QPBO, either exactly or heuristically. These advances in solving QPBO problems are certainly the major driving force behind the recent surge of papers on quadratization techniques. In particular, quadratization has emerged as one of the most successful approaches to the solution of very large-scale PBO models arising in computer vision applications.

The objective of the present paper is to initiate a systematic study of quadratizations of pseudo-Boolean functions. Indeed, most previous papers on this topic have concentrated on proposing specific quadratization procedures, on experimentally testing the quadratic optimization models produced by various procedures, or on investigating quadratizations of submodular pseudo-Boolean functions. But our understanding of quadratizations and of their structural properties remains extremely limited at this time. Therefore, we propose in this paper to adopt a more global perspective, and to investigate some of the properties of the class of *all* quadratizations of a given function.

The paper addresses two main types of questions. First, we want to determine the minimum number of auxiliary y-variables required *in the worst case* in a quadratization of an arbitrary function f, as a function of the number of variables of f. This question is rather natural since the complexity of minimizing the quadratic function $g(x, y)$ heavily depends (among other factors) on the number of binary variables (x, y). In Section 3, we establish a lower bound on the number of auxiliary variables required in *any* quadratization of some pseudo-Boolean functions. In Section 4, we establish an upper bound on the required number of variables, of the same order of magnitude as the lower bound. We further investigate similar lower and upper bounds for pseudo-Boolean functions of bounded degree, and for a restricted class of quadratizations. A companion paper (Anthony et al. [1]) determines more precisely the number of auxiliary variables required by quadratizations of symmetric functions.

In Section 5, we provide a complete description of the simplest quadratizations of a single monomial, namely, quadratizations involving only one auxiliary y-variable. Although this question may appear to be rather narrow, it turns out to be more complex than one may expect at first sight. We conclude the paper with a short list of open questions.

2 Main concepts and literature review

Let n be a finite number, $[n] = \{1, 2, \ldots, n\}$, and $2^{[n]} = \{S : S \subseteq [n]\}$. We denote by F_n the set of pseudo-Boolean functions on $\{0,1\}^n$, and by $BF_n \subset F_n$ the set of *Boolean functions* on $\{0,1\}^n$, meaning pseudo-Boolean functions with values in $\{0,1\}$. Note that a pseudo-Boolean function can equivalently be viewed as a *set function* defined on $2^{[n]}$.

As mentioned earlier, every pseudo-Boolean function can be uniquely represented as a multilinear polynomial in its variables (see, e.g., [13, 27]): for every function f on $\{0,1\}^n$, there exists a unique mapping $a\colon 2^{[n]} \to \mathbb{R} : S \mapsto a_S$ such that

$$f(x_1, x_2, \ldots, x_n) = \sum_{S \in 2^{[n]}} a_S \prod_{i \in S} x_i. \tag{3}$$

In the remainder of this paper, we assume that the input of a PBO problem is a pseudo-Boolean function given by its polynomial expression (3), where only the nonzero coefficients a_S are explicitly listed. We define the *degree* of a function as the degree of this unique polynomial. Sometimes, we also consider expressions involving the (Boolean) *complement* $\overline{x} = 1 - x$ of certain variables x. (Note that many applications of PBO naturally arise in this form; when this is not the case, transforming an arbitrary expression of f into its polynomial expression may be computationally expensive.)

Pseudo-Boolean optimization finds its roots in early papers by Fortet [19, 20] and Maghout [38, 39], among others, and was popularized by the monograph Hammer and Rudeanu [27]. Overviews can be found in Boros and Hammer [4], Crama and Hammer [12, 13]. In the 60's and 70's, several authors proposed to handle the PBO problem (1) by reformulating it as an integer linear programming problem, as follows:

1. in the objective function (3), replace each nonlinear monomial $\prod_{i \in S} x_i$ by a new variable y_S, and
2. set up linear constraints forcing $y_S = \prod_{i \in S} x_i$ in all optimal solutions of the resulting 0–1 LP.

Such *linearization* techniques have given rise to an enormous amount of literature; for a (partial) overview see, e.g., the above references, or Burer and Letchford [10], Hansen, Jaumard and Mathon [28], Sherali and Adams [45], etc.

As mentioned in the Introduction, Rosenberg [42] showed that the general PBO problem (1) can also be efficiently reduced to the quadratic case. More precisely, Rosenberg's procedure works as follows:

1. select two variables, say x_i, x_j such that the product $x_i x_j$ appears in a monomial of degree at least 3 in the objective function f;
2. let h_{ij} be the function obtained upon replacing each occurence of the product $x_i x_j$ by a new variable y_{ij} in f;
3. let $g_{ij} = h_{ij} + M(x_i x_j - 2x_i y_{ij} - 2x_j y_{ij} + 3y_{ij})$, where M is a large enough positive number.

It is easy to verify that for each value of x, the minimum of g_{ij} over the auxiliary variable y_{ij} is exactly equal to $f(x)$: indeed, the minimizer is $y^*_{ij} = x_i x_j$, and the penalty term vanishes for this value. The same step can be repeated until the degree of the resulting polynomial is equal to 2, that is, until a quadratization of f is obtained.

Rosenberg's result establishes that every pseudo-Boolean function has a quadratization that can be computed in polynomial time. Perhaps for lack of efficient quadratic optimization algorithms, his approach, however, did not lead to practical applications for about 30 years. Meanwhile, much progress has been done on solving QPBO: depending on the structure of the objective function (e.g., on its density), instances of QPBO involving about 100-200 variables are now frequently solved to optimality, whereas heuristic algorithms provide solutions of excellent quality for instances with up to a few thousand variables. Among recent contributions on QPBO, we can mention for example the experimental work with exact algorithms by Billionnet and Elloumi [2], Boros, Hammer, Sun and Tavares [5], Hansen and Meyer [29], Helmberg and Rendl [30], and with heuristic algorithms by Boros, Hammer and Tavares [6], Glover, Alidaee, Rego and Kochenberger [24], Glover and Hao [25], Merz and Freisleben [40].

Interestingly, much algorithmic progresss on QPBO has been due to the computer vision research community, where quadratic pseudo-Boolean functions (often dubbed "energy functions" in this framework) have proved successful in modelling specific applications such as image restoration or scene reconstruction; see, e.g., Boykov, Veksler and Zabih [7], Kolmogorov and Rother [36], Kolmogorov and Zabih [37], Rother, Kolmogorov, Lempitsky and Szummer [44]. Fast QPBO heuristics building on the roof-duality framework initially introduced by Hammer, Hansen and Simeone [26] have been developed by these researchers (see also Boros et al. [5]) and can efficiently handle the very large-scale, specially structured, sparse instances arising in computer vision applications.

In recent years, the same community has shifted its attention to more complex models, where the "energy function" to be minimized is a higher-degree pseudo-Boolean function. These models are very large, with as many as 10^6 variables and 10^7 terms in their polynomial representation. Traditional approaches based on integer linear formulations would require a similarly high number of variables and constraints, making them computationally unattractive. In this context, several researchers have met considerable success with approaches based on generating a quadratization $g(x, y)$ of the objective function $f(x)$, and on minimizing g instead of f; see, e.g., Boros and Gruber [3], Fix, Gruber, Boros and Zabih [17, 18], Freedman and Drineas [21], Ishikawa [32], Kappes et al. [34], Ramalingam, Russell, Ladický and Torr [41], Rother, Kohli, Feng and Jia [43], etc.

In particular, various *termwise quadratization procedures* based on the following scheme have been proposed. For a real number c, let $sign(c) = +1$ (resp., -1) if $c \geq 0$ (resp., $c < 0$). Then, given $f(x)$ as in Equation (3),

1. for each $S \in 2^{[n]}$, let $g_S(x, y_S)$ be a quadratization of the monomial $sign(a_S) \prod_{i \in S} x_i$, where $(y_S, S \in 2^{[n]})$ are disjoint vectors of auxiliary variables (one vector for each S);
2. let $g(x, y) = \sum_{S \in 2^{[n]}} |a_S| \, g_S(x, y_S)$.

Then, $g(x, y)$ is a quadratization of $f(x)$. Various choices of the subfunctions $g_S(x, y_S)$ can be found in the literature. When a_S is negative, Freedman and Drineas [21] suggest to use the *standard quadratization* $g_S(x,y) = (|S| - 1)y - \sum_{i \in S} x_i y$, where y is a single auxiliary variable (see Section 5 hereunder). More generally, one can observe that for all $S, T \in 2^{[n]}$ with $S \cap T = \emptyset$,

$$-\left(\prod_{i \in S} x_i\right)\left(\prod_{j \in T} \overline{x}_j\right) = \min_{y \in \{0,1\}} \left\{ \left(\sum_{i \in S} \overline{x}_i + \sum_{j \in T} x_j - 1\right) y \right\}. \qquad (4)$$

When a_S is positive, the choice of an appropriate function g_S is less obvious. One possibility is to note that if $j \in S$, then $\prod_{i \in S} x_i = -\overline{x}_j \prod_{i \in S \setminus j} x_i + \prod_{i \in S \setminus j} x_i$. The first term of this decomposition can be quadratized as in Equation (4), and the second term can be handled recursively. This results in a quadratization $g_S(x,y)$ using $|S| - 2$ auxiliary variables for each positive monomial. (Rosenberg's procedure would produce the same number of auxiliary variables.) Ishikawa [32] showed that positive monomials can actually be quadratized using $\left\lfloor \frac{|S|-1}{2} \right\rfloor$ auxiliary variables, and this is currently the best available bound for positive monomials; see also [1, 17, 32]. It may be worth stressing that all these procedures can be implemented in polynomial time, and that the number of auxiliary variables they introduce is at most n times the size of the function f.

In our paper we consider all possible quadratizations of a function, not only the ones obtainable via one of the above mentioned specific procedures. To formalize this, let us associate with a function $f(x)$ the quantity

$$\xi(f) = \min\{m \mid \text{there is a quadratization } g(x,y) \text{ of } f \text{ with } y \in \{0,1\}^m\},$$

that is the minimum number of auxiliary variables one needs to be able to quadratize f. We also define

$$\xi(n,d) \;=\; \max\{\xi(f) \mid f : \{0,1\}^n \to \mathbb{R} \text{ is of degree at most } d\},$$

that is the number of auxiliary variables one may need in the worst case to quadratize a function of n variables that has degree at most d. Finally we define $\xi(n) = \xi(n,n)$, that is the minimum number of auxiliary variables one needs to quadratize any function in n variables. In what follows, we provide tight lower and upper bounds for both $\xi(n)$ and $\xi(n,d)$.

Note that many termwise procedures proposed in the literature (in particular, the procedures of Freedman–Drineas and Ishikawa) yield special types of quadratizations, namely, quadratizations without products of auxiliary variables.

Definition 2 *A quadratic pseudo-Boolean function $g(x, y)$ on $\{0, 1\}^{n+m}$ is called* y-linear *if its polynomial representation does not contain monomials of the form $y_i y_j$ for $i, j \in [m]$, $i \neq j$.*

When $g(x, y)$ is y-linear, it can be written as $g(x, y) = q(x) + \sum_{i=1}^{m} \ell_i(x) y_i$, where $q(x)$ is quadratic in x and each ℓ_i is a linear function of x. Then, when minimizing g over y, each product $\ell_i(x) y_i$ simply takes the value $\min\{0, \ell_i(x)\}$, that is,

$$f(x) = \min_{y \in \{0,1\}^m} g(x, y) = q(x) + \sum_{i=1}^{m} \min\{0, \ell_i(x)\}. \tag{5}$$

Hence, a y-linear quadratization of $f(x)$ produces an alternative representation of f in the x-variables only.

To study this type of quadratizations, we introduce

$$\zeta(f) = \min\{m \mid \text{there is a } y\text{-linear quadratization } g(x, y) \text{ of } f \text{ with } y \in \{0, 1\}^m\},$$

that is the minimum number of auxiliary variables one needs in a y-linear quadratization of f, and we define

$$\zeta(n) \;=\; \max\{\zeta(f) \mid f : \{0, 1\}^n \to \mathbb{R}\},$$

that is the minimum number of auxiliary variables one needs in a y-linear quadratization of a function in n variables, in the worst case. In the sequel we develop lower and upper bounds for $\zeta(n)$, as well.

Buchheim and Rinaldi [8, 9] have developed a very different type of reduction of nonlinear binary optimization problems to the quadratic case. Their approach is polyhedral: essentially, they show that the separation problem for the polyhedron defined by the classical linearization of PBO can be reduced to a separation problem for the quadratic case. The authors also mention in [8] some connections between their approach and Rosenberg's technique, but the relation to the more general quadratization framework discussed in this paper remains to be explored.

Quadratizations have been independently investigated in the constraint satisfaction literature; see, e.g., Živný, Cohen and Jeavons [47]. These authors proved, among other results, that submodular pseudo-Boolean functions of degree 4 or more do not necessarily have submodular quadratizations. Related questions are also investigated in [33, 35, 48].

Finally, although there is some superficial resemblance between the use of auxiliary variables in quadratization procedures and in so-called "extended formulations" of combinatorial optimization problems (see, e.g., Conforti, Cornuéjols and Zambelli [11]), the connection between these two fields of research does not appear to run deeper.

3 Lower bounds on the number of auxiliary variables

Ishikawa [32] observed that, by relying on his procedure for positive monomials and on Freedman and Drineas' procedure for negative monomials (see Section 2), any pseudo-Boolean function can be quadratized using at most $t\left\lfloor \frac{d-1}{2} \right\rfloor$ auxiliary variables, where d is the degree of the polynomial (3) and t is the number of terms with nonzero coefficients. Since t can be as large as $\binom{n}{d}$, this yields a (tight) upper bound $O(n^d)$ on the number of auxiliary variables introduced by Ishikawa's procedure for a polynomial of fixed degree d, and $O(n\,2^n)$ variables for an arbitrary function. The same asymptotic bounds can be derived for the procedures proposed by Rosenberg or by Freedman and Drineas (see Section 2). Note that the bounds are attained for "complete" functions containing all possible positive monomials of degree d and smaller, for each $d \leq n$.

These observations raise the question of determining how many auxiliary variables are required in a quadratization, *independently of the procedure used to produce the quadratization.* The following result provides a partial answer to this question.

Theorem 1 *There are pseudo-Boolean functions of n variables for which* *every quadratization must involve at least $\Omega(2^{n/2})$ auxiliary variables. In particular, we have $\xi(n) \geq 2^{n/2} - n - 1$.*

Proof. Suppose f is a function of $x_1, x_2, \ldots, x_n$. If f can be quadratized using m auxiliary variables $y_1, y_2, \ldots, y_m$, then there is a quadratic pseudo-Boolean function $g(x, y) : \{0,1\}^{n+m} \to \mathbb{R}$ of the form

$$g(x,y) = a + \sum_{i,j=1}^{n} b_{ij} x_i x_j + \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} x_i y_j + \sum_{i,j=1}^{m} d_{ij} y_i y_j$$

(where $x_i x_i = x_i$ and $y_i y_i = y_i$) such that

$$f(x) = \min\{g(x,y) : y \in \{0,1\}^m\} \quad \text{for all } x \in \{0,1\}^n.$$

Equivalently, we can write $f(x) = g(x, y^*(x))$, where

$$y^*(x) = \operatorname{argmin}\{g(x,y) : y \in \{0,1\}^m\}. \tag{6}$$

Equation (6) enables us to view each y_i^* as a Boolean function of x: $y_i^*(x) = h_i(x)$, say. Then,

$$f(x) = a + \sum_{i,j=1}^{n} b_{ij} x_i x_j + \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} x_i h_j(x) + \sum_{i,j=1}^{m} c_{ij} h_i(x) h_j(x).$$

This shows that f is a linear combination of the following set of pseudo-Boolean functions defined on $\{0,1\}^n$, where h denotes $(h_1, \ldots, h_m)$:

$$L_h = \{1, x_i x_j, x_i h_r, h_r h_s : 1 \leq i, j \leq n, 1 \leq r, s \leq m\}.$$

Note that

$$|L_h| = \ell(n,m) = 1 + n + \binom{n}{2} + mn + m + \binom{m}{2}.$$

The set of pseudo-Boolean functions of n variables, F_n, forms a vector space of dimension 2^n isomorphic to $\mathbb{R}^{2^n}$. The set of monomials $\{\prod_{i \in S} x_i : S \in 2^{[n]}\}$ forms a basis of this vector space, as expressed by the unicity of the representation (3).

Let V_m be the set of pseudo-Boolean functions of n variables which can be quadratized using at most m auxiliary variables, and regard V_m as a subset of $\mathbb{R}^{2^n}$. The discussion above shows that for any $f \in V_m$, there exists some choice of Boolean functions $h_1, h_2, \ldots, h_m$, such that f is contained in the subspace $\mathrm{sp}(L_h)$ of $\mathbb{R}^{2^n}$ spanned by the functions in L_h. It follows that V_m is contained in the union $\bigcup_h \mathrm{sp}(L_h)$, where the union is over all possible choices of m Boolean functions $(h_1, h_2, \ldots, h_m)$ of n variables. But there is only a finite number, 2^{2^n}, of possibilities for each of the Boolean functions h_i. So V_m is contained in a finite union of subspaces, each of dimension at most $\ell(n, m)$. If all pseudo-Boolean functions can be quadratized using m auxiliary variables, then V_m — and hence this union — must be the whole space $F_n \cong \mathbb{R}^{2^n}$. That cannot be the case if $\ell(n, m) < 2^n$. In other words, if m auxiliary variables suffice to quadratize any pseudo-Boolean function of n variables, then $\ell(n, m) \geq 2^n$. This implies that $m \geq 2^{n/2} - n - 1$ from which $m = \Omega(2^{n/2})$ follows for $n \geq 6$. □

Let us formulate a few comments about Theorem 1. First, its proof reveals that for *almost all* pseudo-Boolean functions, any quadratization must use at least $\Omega(2^{n/2})$ auxiliary variables (in the sense that the set of those functions which can be quadratized with fewer variables, regarded as a subset of $\mathbb{R}^{2^n}$, has Lebesgue measure zero). In order to avoid misconceptions, we stress that this conclusion should not be interpreted as a "negative result" showing somehow that quadratization inevitably leads to an exponential blowout of the size of the PBO problem and is, therefore, computationally intractable. In fact, as mentioned at the beginning of this section, many quadratization procedures actually have polynomial complexity, as a function of the size of the polynomial representation of f. The exponential lower bound on $\xi(n)$ in Theorem 1 is rather due to the fact that almost all pseudo-Boolean functions have 2^n terms, since F_n has dimension 2^n. (Similar observations would actually apply to classical *linearization* methods.) In this sense, deriving an exponential lower bound on the required number of auxiliary variables is not, in itself, a very surprising result. Our interest, however, is in the *exact* value of this bound, and more specifically, in the fact that it grows like $2^{n/2}$. We return to these comments in Section 4.2.

The proof of Theorem 1 also shows that if we want to write a pseudo-Boolean function in the form $f = \sum_{i,j=1}^{m} a_{ij} h_i h_j$, where $h_1, \ldots, h_m$ are arbitrary Boolean functions, then m must be $\Omega(2^{n/2})$ for some functions. This appears to be a stronger result than Theorem 1, since it does not explicitly refer to quadratizations.

Let us now turn to some extensions of Theorem 1, where we further specify either the class of pseudo-Boolean functions, or the class of quadratizations that we consider. We first analyze the case of pseudo-Boolean functions of bounded degree d, that is, functions expressed by polynomials of degree d.

Theorem 2 *For each fixed d, there are pseudo-Boolean functions of n variables and of degree d for which every quadratization must involve at least $\Omega(n^{d/2})$ auxiliary variables. In particular, we have $\xi(n,d) \geq \frac{n}{d}^{d/2}$, if n is large enough.*

Proof. The set of pseudo-Boolean functions of n variables and of degree (at most) d, say $F_{n,d}$, is a linear subspace of the space $F_n = F_{n,n}$. The dimension of $F_{n,d}$ is $dim(F_{n,d}) = \sum_{k=0}^{d} \binom{n}{k}$.

So, if all functions of $F_{n,d}$ can be quadratized using m auxiliary variables, then each of the subspaces $\mathrm{sp}(L_h)$ introduced in the proof of Theorem 1 must be of dimension at least $dim(F_{n,d})$, that is, $\ell(n,m) \geq \binom{n}{d}$. This implies that $m = \Omega(n^{d/2})$. □

We show next that y-linear quadratizations (Definition 2) must necessarily contain many auxiliary variables. (Recall that Ishikawa's procedure yields y-linear quadratizations involving $O(n\,2^n)$ auxiliary variables.)

Theorem 3 *There are pseudo-Boolean functions of n variables for which every y-linear quadratization must involve at least $\Omega(2^n/n)$ auxiliary variables. In particular, we have $\zeta(n) \geq \frac{2^n}{n+1} - \frac{n+1}{2}$.*

Proof. Let W_m be the set of pseudo-Boolean functions for which there is a y-linear quadratization involving at most m auxiliary variables. We can repeat the argument given in the proof of Theorem 1, omitting the products $h_i h_j$ from L_h when $i \neq j$, to obtain a set L'_h, of size $\ell'(n,m) = 1 + n + \binom{n}{2} + mn + m$. We conclude as before that for all pseudo-Boolean functions to be quadratizable in this way, we would need $W_m = \mathbb{R}^{2^n}$, and so $\ell'(n,m) \geq 2^n$, from which the claim follows. □

4 Upper bounds on the number of auxiliary variables

The remarks formulated after the proof of Theorem 1 seem to suggest that the lower bound stated in this theorem may not be very strong. Therefore, we would like to derive now some upper bounds on the number of auxiliary variables required in a (best possible) quadratization.

4.1 Minterm quadratization

We start with a simple observation.

Proposition 1 *Let $f(x) \in F_n$ be a pseudo-Boolean function on $\{0,1\}^n$ and let M be an arbitrary upper bound on f (e.g., the sum of the positive coefficients in the multilinear expression of f). Then*

$$g(x,y) = M + \sum_{u \in \{0,1\}^n} (M - f(u)) \left(\sum_{u_i = 1} \overline{x}_i + \sum_{u_j = 0} x_j - 1 \right) y_u$$

is a y-linear quadratization of f involving 2^n auxiliary variables.

Proof. It is easy to check directly the validity of the statement. We present here a slightly roundabout argument, which highlights the connections with earlier arguments.

Write $f = M + (f - M)$, and consider the so-called "minterm normal form" (see [13])

$$f(x) = M + \sum_{u \in \{0,1\}^n} (f(u) - M) \left(\prod_{u_i = 1} x_i \right) \left(\prod_{u_j = 0} \overline{x}_j \right).$$

All coefficients $f(u) - M$ in this expression are negative or null. Therefore, all monomials can be quadratized as in Equation (4), and this yields the expression $g(x, y)$ in the statement. □

In spite of its simplicity, the upper bound on the number of auxiliary variables given in Proposition 1 is already stronger than the $O(n2^n)$ bound derived, e.g., from Rosenberg's or Ishikawa's work. An easy generalization goes as follows.

Proposition 2 *Assume that the sets $B_k = \{x \in \{0,1\}^n : x_i \ = 1 \text{ for } i \in P_k, x_j \ = 0 \text{ for } j \in N_k\}$, $k = 1, \ldots, m$, define a partition of $\{0,1\}^n$ into m subcubes such that f takes constant value f_k on each B_k. Then,*

$$g(x, y) = M + \sum_{k=1}^{m} (M - f_k) \left(\sum_{i \in P_k} \overline{x}_i + \sum_{j \in N_k} x_j - 1 \right) y_k$$

is a y-linear quadratization of f involving m auxiliary variables.

4.2 Universal sets

In order to improve the upper bound on the number of auxiliary variables required in a quadratization, we need to introduce a few definitions.

Definition 3 *Let $\mathcal{P} \subseteq F_n$ be a subset of pseudo-Boolean functions and let $\mathcal{U} \subseteq BF_n$ be a subset of Boolean functions. We say that $\mathcal{U}$ is a* universal set *for $\mathcal{P}$ if, for every function $f \in \mathcal{P}$, there is a quadratization $g(x, y)$*

of f requiring $m \leq |\mathcal{U}|$ auxiliary variables $y_1, \ldots, y_m$ and there is a subset $\{y_1^*, \ldots, y_m^*\} \subseteq \mathcal{U}$ such that $y^*(x) = (y_1^*(x), \ldots, y_m^*(x))$ is a minimizer of $g(x, y)$ for all $x \in \{0,1\}^n$ (as in Equation (6)).

The main clause in Definition 3 is that, when $\mathcal{U}$ is a universal set, then all minimizers $(y_1^*(x), \ldots, y_m^*(x))$ can be chosen in $\mathcal{U}$, for all functions in $\mathcal{P}$. Clearly, BF_n itself is a universal set for every set of functions $\mathcal{P}$, and it is not obvious that there should be smaller ones when $\mathcal{P} = F_n$. We are now going to show, however, that rather small universal sets for F_n can be constructed by relying on the concept of *pairwise covers*.

Definition 4 *When $\mathcal{F}, \mathcal{H} \subseteq 2^{[n]}$ are two hypergraphs, we say that $\mathcal{H}$ is a* pairwise cover *of $\mathcal{F}$ if, for every set $S \in \mathcal{F}$ with $|S| \geq 3$, there are two sets $A(S), B(S) \in \mathcal{H}$ such that $|A(S)| < |S|$, $|B(S)| < |S|$, and $A(S) \cup B(S) = S$.*

We can now state:

Theorem 4 *If $\mathcal{F}, \mathcal{H} \subseteq 2^{[n]}$ are two hypergraphs such that $\mathcal{H} \subseteq \mathcal{F}$ and $\mathcal{H}$ is a pairwise cover of $\mathcal{F}$, then $\mathcal{U}(\mathcal{H}) = \left\{ \prod_{j \in H} x_j : H \in \mathcal{H} \right\}$ is a universal set for the set of pseudo-Boolean functions of the form $f(x) = \sum_{S \in \mathcal{F}} a_S \prod_{j \in S} x_j$, and hence every function of this form has a quadratization using at most $|\mathcal{H}|$ auxiliary variables.*

We will use several special cases of this result, e.g., when $\mathcal{F} = 2^{[n]}$ and f is an arbitrary function, or when $\mathcal{F}$ contains all subsets of size at most d and f is a fixed degree-d polynomial. For now, we start with a proof of the proposition.

Intuitively, in this proof, the pairwise cover $\mathcal{H}$ will tell us how to partition each monomial of the form $\prod_{j \in S} x_j$ into a product of two monomials $\left(\prod_{j \in A(S)} x_j\right) \left(\prod_{j \in B(S)} x_j\right)$, which will be subsequently replaced by a product of two auxiliary variables $y_{A(S)}\, y_{B(S)}$.

Proof. Let $|\mathcal{H}| = m$ and consider a function $f(x) = \sum_{S\in\mathcal{F}} a_S \prod_{j\in S} x_j$. Note that for every choice of nonnegative coefficients b_S, $S \in \mathcal{F}$, we have

$$f(x) \;=\; \min_{y\in\{0,1\}^m} \sum_{S\in\mathcal{F}} a_S \prod_{j\in S} x_j + \sum_{H\in\mathcal{H}} b_H \left(y_H \left(|H| - \frac{1}{2} - \sum_{j\in H} x_j \right) + \frac{1}{2} \prod_{j\in H} x_j \right) \tag{7}$$

for all $x \in \{0,1\}^n$. This is because $y^*_H = \prod_{j\in H} x_j$ minimizes the right-hand side of (7) for all x, and for this value the second summation in the right-hand side is identically zero. (This reflects the fact that $y_H \left(|H| - \frac{1}{2} - \sum_{j\in H} x_j \right)$ is nothing but a variant of the quadratization in Equation (4) for the negative monomial $-\frac{1}{2} \prod_{j\in H} x_j$.)

We now specify the coefficients b_H as follows: For $H \in \mathcal{F}$, $H \notin \mathcal{H}$, we let $b_H = 0$. For $H \in \mathcal{H}$, we let

$$\frac{1}{2} b_H \;=\; \sum_{\substack{S\in\mathcal{F}:\\ H\in\{A(S),B(S)\}}} \left(|a_S| + \frac{1}{2} b_S \right). \tag{8}$$

Note that for each $H \in \mathcal{H}$, the right-hand side of Equation (8) only involves subsets S with $|S| > |H|$. Thus the system of equations (8), for $H \in \mathcal{H}$, is triangular and has a nonnegative feasible solution $b_S \geq 0$ for all $S \in \mathcal{F}$.

Let us substitute this solution in Equation (7), and let us finally replace every occurence of a term $\prod_{j\in T} x_j$ in Equation (7) by $y_{A(T)} y_{B(T)}$. Note that this construction is well defined since $\mathcal{H} \subseteq \mathcal{F}$. It yields a quadratic function $g(x, y)$. We claim that $g(x, y)$ is a quadratization of $f(x)$.

More precisely, consider a point $x \in \{0,1\}^n$. We are going to show that, here again, $y^*_H = \prod_{j\in H} x_j$, for all $H \in \mathcal{H}$, minimizes $g(x, y)$, which entails that $f(x) = \min_{y\in\{0,1\}^m} g(x, y)$ and that $\mathcal{U}(\mathcal{H})$ is a universal set.

To see this, consider an arbitrary set $H \in \mathcal{H}$ and write $g(x, y) = c(x, y) y_H + d(x, y)$, where $c(x, y)$ and $d(x, y)$ do not depend on y_H and are uniquely defined by this condition. More precisely, when $H \in \{A(S), B(S)\}$, define

$R(S)$ to be such that $\{H, R(S)\} = \{A(S), B(S)\}$. Then,

$$\begin{aligned} c(x,y) &= \sum_{\substack{S\in\mathcal{F}:\\ H\in\{A(S),B(S)\}}} a_S\, y_{R(S)} \;+\; b_H \left(|H| - \frac{1}{2} - \sum_{j\in H} x_j\right) \;+ \frac{1}{2} \sum_{\substack{S\in\mathcal{H}:\\ H\in\{A(S),B(S)\}}} b_S\, y_{R(S)} \\ &= \sum_{\substack{S\in\mathcal{F}:\\ H\in\{A(S),B(S)\}}} (a_S + \frac{1}{2} b_S)\, y_{R(S)} \;+\; b_H \left(|H| - \frac{1}{2} - \sum_{j\in H} x_j\right). \qquad (9) \end{aligned}$$

If $\prod_{j\in H} x_j = 1$, then we get

$$\begin{aligned} c(x,y) &= \sum_{\substack{S\in\mathcal{F}:\\ H\in\{A(S),B(S)\}}} (a_S + \frac{1}{2} b_S)\, y_{R(S)} - \frac{1}{2} b_H \qquad (10) \\ &\leq \sum_{\substack{S\in\mathcal{F}:\\ H\in\{A(S),B(S)\}}} \left(|a_S| + \frac{1}{2} b_S\right) - \frac{1}{2} b_H \qquad (11) \\ &= 0 \end{aligned}$$

where the last equality is implied by Equation (8). Thus, $c(x,y) \leq 0$ and hence $y^*_H = 1$ minimizes $g(x,y)$.

If $\prod_{j\in H} x_j = 0$, then $\sum_{j\in H} x_j \leq |H| - 1$, and thus we get by (9)

$$\begin{aligned} c(x,y) &\geq \sum_{\substack{S\in\mathcal{F}:\\ H\in\{A(S),B(S)\}}} (a_S + \frac{1}{2} b_S)\, y_{R(S)} + \frac{1}{2} b_H \qquad (12) \\ &\geq \frac{1}{2} b_H - \sum_{\substack{S\in\mathcal{F}:\\ H\in\{A(S),B(S)\}}} \left(|a_S| + \frac{1}{2} b_S\right) \qquad (13) \\ &= 0. \end{aligned}$$

Here the first inequality is implied by $b_H \geq 0$, the second follows from the inequalities $(a_S + \frac{1}{2} b_S)\, y_{R(S)} \geq \left(-|a_S| - \frac{1}{2} b_S\right)$, while the last equality follows by (8). Thus, $c(x,y) \geq 0$ implies that $y^*_H = 0$ minimizes $g(x,y)$. □

It may be interesting to stress that the procedure described in the proof of Theorem 4 is entirely constructive and can be performed efficiently. Indeed, observe first that, for any function f of the form given in the theorem, one can easily extend the set of terms of f to a larger set $\mathcal{F}^*$ such that $\mathcal{F}^*$ is a pairwise cover of itself and $|\mathcal{F}^*| \leq n|\mathcal{F}|$: for each set $S \in \mathcal{F}$, say for example $S = \{1, 2, \ldots, s\}$ with s even, it suffices to include in $\mathcal{F}^*$ all $(s-1)$ subsets

$$S, \{1,2\}, \{3, \ldots, s\}, \{3,4\}, \{5, \ldots, s\}, \{5,6\}, \{7, \ldots, s\}, \ldots, \{s-1, s\}.$$

The resulting hypergraph $\mathcal{F}^*$ is a pairwise cover since in this list, every subset of cardinality at least three is the union of the next two subsets. Moreover, the function f can be rewritten as $f(x) = \sum_{S \in \mathcal{F}^*} a_S \prod_{j \in S} x_j$, with $a_S = 0$ when $S \notin \mathcal{F}$. It follows that the construction described in the proof of Theorem 4 can be used (with $\mathcal{F} = \mathcal{H} = \mathcal{F}^*$) to quadratize in polynomial time any pseudo-Boolean function while only using $O(n|\mathcal{F}|)$ auxiliary variables.

Remark 1 From a practical point of view, there may be more effective ways of producing a "small" hypergraph $\mathcal{F}^*$ which extends $\mathcal{F}$ and which is a pairwise cover of itself. For instance, in their reduction of PBO to the quadratic case, Buchheim and Rinaldi [8] also use a transformation which replaces each monomial of f by a product of two lower-degree monomials. They say that $\mathcal{F}$ is *reducible* when (in our terminology) $\mathcal{F}$ is a pairwise cover of itself, and they describe a heuristic algorithm to obtain a small reducible extension of $\mathcal{F}$. They observe that finding the smallest reducible extension of $\mathcal{F}$ is equivalent to finding the sequence of substitutions which introduces the smallest possible number of auxiliary variables in Rosenberg's procedure. Boros and Hammer [4] discuss the latter problem and show that it is NP-hard. □

Of course, the $O(n|\mathcal{F}|)$ upper bound that we have derived from Theorem 4 is not very original, since the same bound is also achieved by any of the termwise procedures already described in previous sections. However, the use of pairwise covers allows us to reach beyond the grasp of termwise procedures. (Fix et al. [18] provide experimental evidence for the merits of non-termwise quadratization procedures.) In particular, we are now ready to establish a tight upper bound on the number of auxiliary variables required in a quadratization of an arbitrary pseudo-Boolean function.

Theorem 5 *Every pseudo-Boolean function of n variables has a quadratization involving at most $\xi(n) \leq 2^{\lceil n/2 \rceil} + 2^{\lfloor n/2 \rfloor} - 2 = O(2^{n/2})$ auxiliary variables.*

Proof. Let $\mathcal{H}^{oe}$ contain all nonempty subsets of $[n]$ consisting either only of odd integers, or only of even integers. Then, $\mathcal{H}^{oe}$ is a pairwise cover of $\mathcal{F} = 2^{[n]}$ with size $|\mathcal{H}^{oe}| = 2^{\lceil n/2 \rceil} + 2^{\lfloor n/2 \rfloor} - 2$. Hence, Theorem 4 implies that every pseudo-Boolean function on $\{0,1\}^n$ has a quadratization using at most $|\mathcal{H}^{oe}|$ auxiliary variables. □

The order of magnitude of the upper bound in Theorem 5 nicely matches the lower bound in Theorem 1 which turns out, therefore, to be stronger than expected. Note that, as already observed after the proof of Theorem 1, the multilinear representation of almost all pseudo-Boolean functions involves 2^n terms. Thus, Theorem 5 is a rather positive result: it shows that the number of auxiliary variables needed in an "optimal" quadratization is usually much smaller than the number of terms of the function. This implies, in particular, that termwise quadratization procedures are wasteful in their use of auxiliary variables, since they would typically require $\Theta(n2^n)$ auxiliary variables for almost all functions, instead of $O(2^{n/2})$ auxiliary variables as in Theorem 5.

Remark 2 Pairwise covers are closely related to hypergraphs called 2-*bases* by Füredi and Katona [23], Frein, Lévêque and Sebő [22], Ellis and Sudakov [15], the only difference being that the subsets $A(S), B(S)$ are not required to be *strict* subsets of S in a 2-base. In particular, $\mathcal{H}^{oe}$ is a 2-base of $2^{[n]}$. The role of odd and even integers in its construction could be replaced by any partition of $[n]$ into two sets V_1, V_2 of nearly-equal sizes $\lceil \frac{n}{2} \rceil$ and $\lfloor \frac{n}{2} \rfloor$. According to Füredi and Katona [23], Erdős has conjectured that this generic construction yields the smallest possible 2-bases of $2^{[n]}$. □

We now establish an upper bound which coincides with the lower bound obtained in Theorem 2 for the bounded-degree case.

Theorem 6 *For each fixed d, every pseudo-Boolean function of n variables and of degree d has a quadratization involving at most $\xi(n,d) = O(n^{d/2})$ auxiliary variables.*

Proof. For any fixed value of d, let $\mathcal{F} = [n]^d = \{S \subseteq [n] : |S| \leq d\}$. In order to establish the theorem, we just need to produce a small pairwise cover of $[n]^d$. For simplicity of the presentation, assume that d is a power of 2, and let $\mathcal{H}^d$ contain all subsets of $[n]$ of sizes $d/2, d/4, d/8, \ldots, 2$. Then, it is easy to see that $\mathcal{H}^d$ is a pairwise cover of $[n]^d$ with size $O(n^{d/2})$. □

Here again, it is interesting to remark that almost all degree-d functions have $\Omega(n^d)$ terms, and that the number of auxiliary variables introduced by the construction of Theorem 6 is only the square root of this number of terms.

Example 1 *An arbitrary (say, random) function of degree 8 has* $\Theta(n^8)$ *terms, and all termwise quadratization methods (that is, almost all known methods) would introduce* $\Omega(n^8)$ *auxiliary variables. By contrast, using sets of sizes 4 and 2, as in Theorem 6, we can obtain a quadratization with only* $O(n^4)$ *new variables and in view of Theorem 2, this is in fact typically best possible.* □

Finally, for the case of y-linear quadratizations, we have the following result, which matches the lower bound of Theorem 3 up to a $\log n$ factor.

Theorem 7 *Every pseudo-Boolean function of* n *variables has a* y*-linear quadratization involving at most* $\zeta(n) = O(\frac{2^n}{n} \log n)$ *auxiliary variables.*

Our proof of this theorem is quite long and is very different from the previous ones. For the sake of readability, we handle it in a separate subsection.

4.3 Proof of Theorem 7

We start by introducing some useful notation. For integers m and n, we let $[m..n] = \{m, m+1, \ldots, n-1, n\}$. So, $[m..n] = \emptyset$ if $m > n$, and $[n] = [1..n]$.

The unit vector with i-th entry equal to 1 and all other entries equal to 0 is denoted by e^i. The all-zero vector is denoted by $\mathbf{0}$ (boldface zero).

For $a \in \{0,1\}^n$, the *Hamming weight* (or 1-norm) of a is given by $|a| := \sum_{i=1}^n a_i$. The k-th *layer* of $\{0,1\}^n$, with $k \in [0..n]$, is $Q_k := \{a \in \{0,1\}^n : |a| = k\}$. It has $|Q_k| = \binom{n}{k}$ elements. For $a \in Q_k$ and $k \in [0..n-1]$, the *upper neighborhood of* a is the set $N(a) := \{b \in Q_{k+1} : |b-a| = 1\}$. Each element in $N(a)$ is an *upper neighbor* of a.

Definition 5 *Let $\mathcal{A}^n := \{A_0, A_1, \ldots, A_{n-1}\}$ be a family of sets such that $\emptyset \neq A_k \subseteq Q_k$ for all $k \in [0..n-1]$, and let $\mathcal{D} = \{\Delta(a) : a \in \bigcup_{k=0}^{n-1} A_k\}$ be a family of sets such that $\emptyset \neq \Delta(a) \subseteq N(a)$ for all a. We say that $\mathcal{D}$ is an* attractive partition of $\{0,1\}^n$ induced by $\mathcal{A}^n$ *if*

$$\bigcup_{a \in A_k} \Delta(a) = Q_{k+1} \quad \textit{and} \quad \Delta(a) \cap \Delta(a') = \emptyset,$$

for all $k \in [0..n-1]$ and for all $a, a' \in A_k$ with $a \neq a'$. We say that $\mathcal{A}^n$ is an inductor *of the partition. Its size is defined as $|\mathcal{A}^n| := \sum_{k \in [0..n-1]} |A_k|$.*

Note that $A_0 = \{\mathbf{0}\}$ by definition. Strictly speaking, $\mathcal{D} \cup \{\mathbf{0}\}$ is a partition of $\{0,1\}^n$, rather than $\mathcal{D}$ itself. Also, even though $\mathcal{A}^n$ does not uniquely define $\mathcal{D}$, we frequently find it convenient to identify the partition with its inductor when this does not create confusion.

For each element $a \in \bigcup_{k=0}^{n-1} A_k$ of an inductor, let us define the set

$$\delta(a) := \{i \in [1..n] : a_i = 0 \text{ and } a + e^i \in \Delta(a)\}.$$

Let also $\tilde{a}$ denote the binary vector

$$\tilde{a}_i := \begin{cases} a_i & \text{if } i \notin \delta(a), \\ 1 & \text{otherwise,} \end{cases}$$

and define the *subcube of* $\{0,1\}^n$ *induced by* a and $\Delta(a)$ as

$$[a, \tilde{a}] = \{b \in \{0,1\}^n : a \leq b \leq \tilde{a}\}$$

(this is the smallest subcube that contains a and all vectors in $\Delta(a)$).

Given a pseudo-Boolean function $f : \{0,1\}^n \to \mathbb{R}$, we can use an attractive partition to construct a y-linear quadratization of f. More specifically, we will construct a sequence of y-linear quadratic pseudo-Boolean functions $g_0, g_1, \ldots, g_n : \{0,1\}^{n+m} \to \mathbb{R}$ (m to be estimated) such that g_{k+1} results from a local "adjustment" of g_k on layer Q_k. Each function g_k, when minimized on the auxiliary variables $y \in \{0,1\}^m$, produces a function $\sigma_k(x) = \min_{y\in\{0,1\}^m} g_k(x,y)$ that bounds f in the following way: $\sigma_k(x) = f(x)$ for all $x \in \{0,1\}^n$ such that $|x| \leq k$, and $\sigma_k(x) \geq f(x)$ whenever $|x| > k$. In particular, g_n is a y-linear quadratization of f.

To provide further intuition, let us first recall that the auxiliary variables in a y-linear quadratization appear in terms like $\ell_i(x)y_i$ where ℓ_i is a linear functions of the original variables. When minimizing over y_i, this term contributes the nonpositive quantity $\min\{0, \ell_i(x)\}$ to the value of f; see Equation (5). Our plan is to use a series of y-linear adjustments such that, starting from a majorant σ_0 of f, we can reduce it to f, layer after layer. Unfortunately, when we introduce the new variables to adjust σ_k to the values of f in layer Q_k, we may decrease the values of our new approximation for vectors in $Q_{k+1}, \ldots, Q_n$ by much more than intended. To make sure that the sequence remains above f, therefore, we will start with a symmetric majorant of f, which is increasingly larger than f on higher layers, to preventively compensate for later "accidental" decreases.

Accordingly, our first ingredients in the construction of a y-linear quadratization are symmetric pseudo-Boolean functions $s_k : \{0,1\}^n \to \mathbb{R}$, for $k = 0, ..., n+1$, in the same variables as f:

$$s_k(x) := \begin{cases} 0 & \text{if } |x| < k \\ D_k & \text{if } |x| \geq k, \end{cases} \tag{14}$$

where $D_k \geq 0$ are constants to be specified later.

According to a result by Anthony et al. [1] about the k-out-of-n function, s_k has a y-linear quadratization, say $\hat{s}_k(x,y)$, requiring only $\lceil \frac{n}{2} \rceil$ auxiliary variables. (Fix [16] established an upper bound of $n-1$ which would be sufficient for our purpose.) We denote by y_j^k, $j = 1, \ldots, \frac{n}{2}$, the auxiliary variables appearing in $\hat{s}_k$, $k = 0, 1, \ldots, n$. We emphasize that for $k \neq \ell$ the functions $\hat{s}_k$ and $\hat{s}_\ell$ depend on disjoint sets of auxiliary variables, and hence

$$\min_y \left(\hat{s}_k(x, y) + \hat{s}_\ell(x, y)\right) \;=\; s_k(x) + s_\ell(x) \tag{15}$$

for all $x \in \{0, 1\}^n$.

We next define the following sequence of quadratic pseudo-Boolean functions: for $k = 0, \ldots, n - 1$,

$$g_0(x, y) := \hat{s}_0(x, y) + \hat{s}_1(x, y), \tag{16}$$

$$g_{k+1}(x, y) := \hat{s}_{k+2}(x, y) + g_k(x, y) + \sum_{a \in A_k} y_a h_a(x), \tag{17}$$

where $y_a \in \{0, 1\}$ is an auxiliary variable for each $a \in A_k$,

$$h_a(x) := \alpha_a \sum_{i \notin \delta(a)} \Big(a_i\, \overline{x}_i + \overline{a}_i\, x_i\Big) - \sum_{i \in \delta(a)} \Big(\sigma_k(a + e^i) - f(a + e^i)\Big) x_i, \tag{18}$$

with $\alpha_a = 1 + \sum_{i \in \delta(a)} |\sigma_k(a + e^i) - f(a + e^i)|$. Finally, for all $k = 0, \ldots, n$,

$$\sigma_k(x) := \min_{y \in \{0,1\}^m} g_k(x, y). \tag{19}$$

At this point, we are ready to specify the constants D_k involved in definition (14) of the functions $s_k(x)$. We set

$$D_0 := f(\mathbf{0}),\; D_1 := \max_{x \in \{0,1\}^n} f(x) - f(\mathbf{0}), \tag{20}$$

and recursively, for $k = 0, \ldots, n - 1$,

$$D_{k+2} := (n - k)\, |A_k| \max_{x \in Q_{k+1}} (\sigma_k(x) - f(x)). \tag{21}$$

Note that σ_k only depends on $D_0, \ldots, D_{k+1}$ (through $g_0, \ldots, g_k$), and hence D_{k+2} is well-defined.

We note some simple consequences of the previous definitions.

Fact 1 *For each* $a \in \bigcup_{k=0}^{n-1} A_k$*,*
(i) $h_a(x) = -\sum_{i \in \delta(a)} \Big(\sigma_k(a+e^i) - f(a+e^i)\Big) x_i$ *when* $x \in [a, \tilde{a}]$*;*
(ii) $h_a(a) = 0$*;*
(iii) $h_a(x) > 0$ *when* $x \notin [a, \tilde{a}]$*;*
(iv) $\min_{y_a} y_a h_a(x) = 0$ *for all* $x \notin [a, \tilde{a}]$ *and for* $x = a$*.*

Proof. If $x \in [a, \tilde{a}]$, then by definition of δ, we have $x_i = a_i$ for all $i \notin \delta(a)$. Hence, the summation

$$\sum_{i \notin \delta(a)} \Big(a_i\, \overline{x}_i + \overline{a}_i\, x_i\Big) \tag{22}$$

vanishes in the definition of $h_a(x)$, and (i) follows.

If $x = a$, then all terms of $h_a(x)$ vanish since $a_i = 0$ when $i \in \delta(a)$. This implies (ii).

If $x \notin [a, \tilde{a}]$, then (22) is positive, and thus (iii) follows by the definition of α_a.

Finally, (iv) is a direct consequence of (ii) and (iii). □

Fact 2 *For each* $k = 0, \ldots, n$*, the function* $g_k(x, y)$ *is* y*-linear. It only depends on the original variables* $x_1, \ldots, x_n$*, on the* $(k+2)\frac{n}{2}$ *auxiliary variables* y_j^ℓ*,* $\ell = 0, 1, \ldots, k+1$*,* $j = 1, \ldots, \frac{n}{2}$ *occuring in* $\hat{s}_0, \ldots, \hat{s}_{k+1}$*, and on the auxiliary variables* y_a *for* $a \in \bigcup_{j \in [0..k-1]} A_j$*.*

Proof. Immediately follows from Equations (16)–(19), by induction. □

In view of Fact 2, the three main terms in (17) depend on disjoint sets of auxiliary variables. Thus,

$$\begin{aligned}\sigma_{k+1}(x) = &\ \min_y g_{k+1}(x,y)\\ = &\ \left(\min_y \hat{s}_{k+2}(x,y)\right) + \left(\min_y g_k(x,y)\right) + \left(\min_y \sum_{a \in A_k} y_a h_a(x)\right)\\ = &\ s_{k+2}(x) + \sigma_k(x) + \left(\min_y \sum_{a \in A_k} y_a h_a(x)\right). \qquad (23)\end{aligned}$$

We will repeatedly rely on equality (23) in the sequel.

We are now ready to establish the main properties of the above construction.

Proposition 3 *If $\mathcal{A}^n$ induces an attractive partition, then for all $x \in \{0,1\}^n$ and all $k \in [0..n]$,*

$$\begin{aligned} f(x) = \sigma_k(x) \quad &\textit{if } |x| \le k,\\ f(x) \le \sigma_k(x) \quad &\textit{if } |x| > k.\end{aligned}$$

In particular, $f(x) = \sigma_n(x)$, and thus $g_n(x,y)$ is a y-linear quadratization of $f(x)$ involving $m = O(n^2) + |\mathcal{A}^n|$ auxiliary variables.

Proof. Let $x \in \{0,1\}^n$ be arbitrary. The proof is by induction on k. In case $k = 0$, (14) and (20) easily imply that, for all $x \in \{0,1\}^n$,

$$\begin{cases} s_0(x) & = f(\mathbf{0}), \text{ and} \\ s_0(x) + s_1(x) & \ge f(x). \end{cases} \qquad (24)$$

In view of (15), (16) and (19), it follows that

$$\begin{aligned} \sigma_0(x) = f(\mathbf{0}) \quad &\text{if } |x| = 0,\\ \sigma_0(x) \ge f(x) \quad &\text{if } |x| > 0.\end{aligned}$$

Now suppose the statement is valid for $k < n$ and let us show that it is also valid for $k+1$. We divide the analysis into three cases.

Case 1: $|x| \leq k$.

Either $x \in A_k$ or, for all $a \in A_k$, $x \notin [a, \tilde{a}]$. In either case, we have $\min_y \sum_{a \in A_k} y_a h_a(x) = 0$ by Fact 1. Furthermore, $s_{k+2}(x) = 0$ by definition, since $|x| < k+2$. Thus, by (23) we get

$$\sigma_{k+1}(x) = \sigma_k(x) = f(x),$$

where the last equality follows by the induction hypothesis.

Case 2: $|x| = k+1$.

Since $\mathcal{A}^n$ is an attractive partition, there are unique $a' \in A_k$ and $i \in \delta(a')$ such that $x = a' + e^i$. Note that for all $a \in A_k \setminus \{a'\}$, we have $x \notin [a, \tilde{a}]$, and hence $\min_{y_a} y_a h_a(x) = 0$ by Fact 1.

Moreover, $y_{a'} h_{a'}(x) = y_{a'}(-\sigma_k(x) + f(x))$ as $x_j = 0$ for all $j \in \delta(a')$ with $j \neq i$. Let us use again relation (23). Since $s_{k+2}(x) = 0$ by definition and since $\sigma_k(x) \geq f(x)$ by our inductive hypothesis, we get

$$\sigma_{k+1}(x) = \sigma_k(x) + \min_{y_{a'} \in \{0,1\}} y_{a'}(-\sigma_k(x) + f(x)) = \sigma_k(x) - \sigma_k(x) + f(x) = f(x).$$

Case 3: $|x| > k+1$.

If, for some $a \in A_k$, $x \notin [a, \tilde{a}]$, then by Fact 1 again

$$\min_y y_a h_a(x) = 0.$$

Thus, we get

$$\min_y \sum_{a \in A_k} y_a h_a(x) = \min_y \sum_{\substack{a \in A_k \\ x \in [a, \tilde{a}]}} y_a \left(- \sum_{i \in \delta(a)} \left(\sigma_k(a + e^i) - f(a + e^i) \right) x_i \right).$$

Note that the multiplier of each variable y_a in the previous expression is nonpositive by our induction hypothesis, and hence

$$\min_y \sum_{a \in A_k} y_a h_a(x) = - \sum_{\substack{a \in A_k \\ x \in [a, \tilde{a}]}} \sum_{i \in \delta(a)} \left(\sigma_k(a + e^i) - f(a + e^i) \right) x_i.$$

Furthermore $s_{k+2}(x) = D_{k+2}$, since $|x| \geq k+2$. Consequently, (23) and the induction hypothesis imply that

$$\begin{aligned} \sigma_{k+1}(x) = & \ D_{k+2} + \sigma_k(x) - \sum_{\substack{a \in A_k \\ x \in [a,\tilde{a}]}} \sum_{i \in \delta(a)} \left(\sigma_k(a+e^i) - f(a+e^i) \right) x_i \\ \geq & \ D_{k+2} + f(x) - \sum_{\substack{a \in A_k \\ x \in [a,\tilde{a}]}} \sum_{i \in \delta(a)} \left(\sigma_k(a+e^i) - f(a+e^i) \right) x_i. \end{aligned}$$

Note that the double summation in this last expression contains at most $(n-k)|A_k|$ terms, each not larger than $\max_{x \in Q_{k+1}}(\sigma_k(x) - f(x))$. Hence

$$\sigma_{k+1}(x) \geq \ D_{k+2} + f(x) - (n-k)\,|A_k| \max_{x \in Q_{k+1}} \left(\sigma_k(x) - f(x)\right),$$

and by (21), $\sigma_{k+1}(x) \geq f(x)$, concluding the inductive argument.

The last assertion of the proposition follows now immediately from Fact 2. □

To complete the proof of Theorem 7, we only need to show that there exists a small enough inductor $\mathcal{A}^n$. We shall derive this from classical combinatorial results related to Turán's problem (see, e.g., Sidorenko [46]).

Definition 6 *A family $\mathcal{T} = \mathcal{T}(n,r,k)$ of k-element subsets of $[n]$ is a* Turán (n,r,k)-system *if every r-element subset of $[n]$ contains at least one element of $\mathcal{T}$. The minimum size of such a family is the* Turán number $T(n,r,k)$.

Turán systems are interesting in our context because they can be used to obtain attractive partitions of $\{0,1\}^n$ by identifying each subset of $[n]$ with its characteristic vector. More specifically, for each k, consider a Turán $(n,k+1,k)$-system $\mathcal{T}_k$ and let A_k be the corresponding subset of Q_k. By Definition 6, for each vector $x \in Q_{k+1}$ (i.e., subset of $[n]$ of size $k+1$), there is a vector

$a(x) \in A_k$ such that $a(x) \leq x$ (if there are several possible choices for $a(x)$, just pick one arbitrarily). Then, for each $a \in A_k$, we can define

$$\Delta(a) = \{x \in Q_{k+1} : a = a(x)\}$$

and this yields an attractive partition induced by $\mathcal{A}^n = \{A_0, \dots, A_{n-1}\}$.

Sidorenko [46] presents procedures to construct Turán $T(n, k+1, k)$-systems, together with the following bound:

$$T(n, k+1, k) \leq \frac{1 + 2\ln k}{k}\binom{n}{k} \qquad \text{for all } k \in [n-1].$$

Since $|A_0| = 1$, we can estimate $|\mathcal{A}^n|$ as

$$\begin{aligned} |\mathcal{A}^n| = \sum_{k=0}^{n-1} |A_k| &\leq 1 + \sum_{k=1}^{n-1} \frac{1 + 2\ln k}{k}\binom{n}{k} \\ &\leq 1 + 2\frac{(1 + 2\ln n)}{n+1} \sum_{k=1}^{n-1} \frac{n+1}{k+1}\binom{n}{k} \\ &\leq 1 + \frac{(1 + 2\ln n)}{n+1} 2^{n+2}. \end{aligned}$$

We conclude that the number of auxiliary variables of g_n in Proposition 3 can be upper bounded as

$$m = O\left(\frac{2^n}{n}\log n\right),$$

and this establishes Theorem 7.

Let us remark that the attractive partition constructed above depends only on the dimension n, and not on the actual pseudo-Boolean function f. Hence, a deeper analysis of the proof of Theorem 7 actually reveals the existence of a *universal set* of Boolean functions of cardinality $O(\frac{2^n}{n}\log n)$ such that any pseudo-Boolean function in n variables has a y-linear quadratization using a subset of this universal set as new variables, in the sense of Definition 3. (This claim holds notwithstanding the fact that all functions s_k, h_a, g_k, depend to some extent on f: the claim is only that the optimal value assumed by the y-variables is independent of f.)

5 Negative monomials

We may see it as an ultimate goal to describe the family of all quadratizations of any arbitrary function f. Although this goal is certainly overly ambitious in general, a more limited objective may be attainable: for instance, we may want to describe *all minimal* quadratizations of certain well-structured functions. As we have mentioned in Section 2 that termwise quadratization procedures play an important role in practical implementations, we concentrate in this section on quadratizations of a single negative monomial. (The paper Anthony et al. [1] investigates the number of auxiliary variables required by quadratizations of symmetric functions.)

Before we turn to this question, we should define more precisely what we mean by describing "all minimal" quadratizations of a function. First, we should note that some quadratizations may be seen as *equivalent*, for most practical purposes, in the sense that they can be transformed into each other by simply *switching* a subset of the auxiliary variables $y_1, \ldots, y_m$, that is, by substituting $\overline{y}_i = 1 - y_i$ for certain variables y_i. We are only interested in describing non-equivalent quadratizations.

Next, although we have focused so far on quadratizations involving the smallest possible number of auxiliary variables (say, *lean* quadratizations), other concepts of minimality may be relevant as well.

Definition 7 *A quadratization $g(x, y)$ of $f(x)$ is* prime *if there is no other quadratization of f, say $h(x, y)$, such that $h(x, y) \leq g(x, y)$ for all $(x, y) \in \{0, 1\}^{n+m}$, and such that $h(x^*, y^*) < g(x^*, y^*)$ for at least one point (x^*, y^*).*

If our objective is to minimize f, then lean or prime quadratizations seem to be especially attractive, although it could be argued that other quadratizations possessing certain structural properties may also be of interest (for instance, if it is easy to compute their minimum).

Let us illustrate these concepts with a couple of examples.

Example 2 *Assume that $g(x, y)$ is a quadratization of $f(x)$ involving m auxiliary variables $(y_1, \ldots, y_m)$, and let y_{m+1} be a new variable. Then, trivially, $g(x, y) + y_{m+1}$ is another quadratization of f which is neither lean nor prime. More generally, if $h(x, y) = \min_{y_k} g(x, y)$ is a quadratization of f for some variable y_k, then $h(x, y) \leq g(x, y)$ for all $(x, y) \in \{0, 1\}^{n+m}$. Hence, here again, g is neither lean nor prime (unless $h(x, y) = g(x, y)$, in which case g does not really depend on y_k).* □

Example 3 *The functions*

$$s_3(x, y) = 2y - x_1y - x_2y - x_3y$$

and

$$s_3^+(x, y) = x_1 + x_2 - x_1y - x_2y + x_3y - x_1x_3 - x_2x_3$$

are two non-equivalent quadratizations of the negative monomial $M_3 = -x_1x_2x_3$ (see Proposition 4 hereunder). The function

$$g(x, y) = 5y - 2x_1y - 2x_2y - 2x_3y$$

is another quadratization of M_3 which also involves a single auxiliary variable. It can be checked that $g(x, y) \geq s_3(x, y)$ and therefore, g is not prime. □

We are now going to extend the previous example and to characterize all lean prime quadratizations of the negative monomials.

Definition 8 *The* standard quadratization *of the negative monomial $M_n = -\prod_{i=1}^n x_i$ is the quadratic function*

$$s_n(x, y) = (n - 1)y - \sum_{i=1}^{n} x_iy. \tag{25}$$

The extended standard quadratization *of M_n is the function*

$$s_n^+(x, y) = (n - 2)x_ny - \sum_{i=1}^{n-1} x_i(y - \overline{x}_n). \tag{26}$$

Let us show that the functions s_n and s_n^+ deserve their names.

Proposition 4 *For all $n \geq 1$, the functions $s_n(x,y)$ and $s_n^+(x,y)$ are quadratizations of $M_n = -\prod_{i=1}^n x_i$.*

Proof. The standard quadratization s_n was already introduced in Section 2 (Equation (4), Freedman and Drineas [21]). For s_n^+, the case $n=1$ is trivial. For $n \geq 2$, fix $x \in \{0,1\}^n$ and suppose first that $x_n = 0$. Then, $M_n(x) = 0$ and $\min_y s_n^+(x,y) = \min_y \sum_{i=1}^{n-1}(1-y)x_i = 0$ is attained for $y=1$. On the other hand, if $x_n = 1$, then $s_n^+(x,y) = s_{n-1}(x,y)$ for all $y \in \{0,1\}$, and the statement follows from the fact that $s_{n-1}(x,y)$ is a quadratization of M_{n-1}. □

When $n \leq 2$, M_n is quadratic and, clearly, it is its own unique prime quadratization. When $n \geq 3$, s_n and s_n^+ are lean quadratizations of M_n since they use a single auxiliary variable (we say that they are 1-*quadratizations*, for short). We claim that, in this case, s_n and s_n^+ are essentially the only prime 1-quadratizations of M_n. More precisely:

Theorem 8 *For $n \geq 3$, assume that $g(x,y)$ is a prime 1-quadratization of M_n. Then, up to an appropriate permutation of the x-variables and up to a possible switch of the y-variable, either $g(x,y) = s_n$ or $g(x,y) = s_n^+$.*

Proof. The proof involves a detailed analysis which turns out to be different according to whether $n=3$ or $n \geq 4$. For the sake of brevity, we restrict ourselves here to the generic case $n \geq 4$ and we refer the reader to the technical report [14] for the special case $n=3$.

So, assume now that $n \geq 4$ and that $g(x,y)$ is a 1-quadratization of M_n. Since $M_n(x) = \min_{y \in \{0,1\}} g(x,y)$ for all binary vectors x, we can assume $g(0,0) = 0$ after substituting $\overline{y}$ for y if necessary. Thus, without any loss of generality we can write

$$g(x,y) \;=\; ay \;+\; \sum_{i=1}^n b_i x_i y \;+\; \sum_{i=1}^n c_i x_i \;+\; \sum_{1 \leq i < j \leq n} p_{ij} x_i x_j. \qquad (27)$$

Let us introduce some useful notations. For any subset $S \subseteq N = [n]$, we write $b(S) = \sum_{i \in S} b_i$, $c(S) = \sum_{i \in S} c_i$ and $p(S) = \sum_{i,j \in S,\ i<j} p_{ij}$. Furthermore, since binary vectors can be viewed as characteristic vectors of subsets, we simply write

$$g(S, y) \ = \ ay \ + \ b(S)y \ + \ c(S) \ + \ p(S)$$

instead of (27), when x is the characteristic vector of S.

Then, the fact that g is a quadratization of M_n can be expressed as

$$0 = \min_{y \in \{0,1\}} \ (a + b(S))y \ + \ c(S) \ + \ p(S) \qquad \text{for all } S \subset N, \tag{28}$$

$$-1 = \min_{y \in \{0,1\}} \ (a + b(N))y \ + \ c(N) \ + \ p(N). \tag{29}$$

Let us now note that by (28), we have $g(0, 1) \geq 0$, and hence

$$a \geq 0. \tag{30}$$

Furthermore, we must have $g(\{i\}, 0) \geq 0$ for all $i \in N$ since $n > 1$, implying

$$c_i \geq 0 \ \text{ for all } i \in N. \tag{31}$$

Let us partition the set of indices as $N = N^0 \cup N^+$, where

$$N^0 = \{u \in N \mid c_u = 0\}, \tag{32}$$

$$N^+ = \{i \in N \mid c_i > 0\}. \tag{33}$$

Since $g(\{i\}, 0) = c_i$, relation (28) implies

$$g(\{i\}, 1) = a + b_i + c_i = 0 \ \text{ for all } i \in N^+, \ \text{ and} \tag{34}$$

$$g(\{u\}, 1) = a + b_u \geq 0 \ \text{ for all } u \in N^0. \tag{35}$$

Let us next write (28) for subsets of size two. Consider first a pair $u, v \in N^0$, $u \neq v$. Since $c_u = c_v = 0$, we get $g(\{u, v\}, y) = (a + b_u + b_v)y + p_{uv}$, implying

$$\min\{p_{uv}, a + b_u + b_v + p_{uv}\} \ = \ 0. \tag{36}$$

Let us consider next $i, j \in N^+$, $i \neq j$. Then, by (34) and by the definitions we get $g(\{i,j\},1) = p_{ij} - a \geq 0$. This, together with (30) implies that $p_{ij} \geq a \geq 0$. Thus, $g(\{i,j\},0) = c_i + c_j + p_{ij} > 0$ implying that $g(\{i,j\},1) = 0$, that is,

$$p_{ij} = a \geq 0 \text{ for all } i, j \in N^+. \tag{37}$$

This allows us to establish a first property of N^0.

Claim 1 $N^0 \neq \emptyset$.

Proof. If $N^0 = \emptyset$, then we have $g(N,y) = (a + b(N^+))y + c(N^+) + \binom{|N^+|}{2}a$ by (37). Since $|N^+|a + b(N^+) + c(N^+) = 0$ by (34), we get $g(N,1) = \binom{|N^+|-1}{2}a \geq 0$ by (30), and $g(N,0) = c(N^+) + \binom{|N^+|}{2}a \geq 0$ by (30) and (31). This contradicts (29) and proves the claim. □

In contrast with Claim 1, the set N^+ may be empty or not.

Claim 2 *If $N^+ = \emptyset$ and $N = N^0$, then $p_{uv} = 0$ for all $u, v \in N$. Furthermore,*

$$\begin{aligned} &a + b(S) \geq 0 \textit{ for all subsets } S \neq N, \textit{ and} \\ &a + b(N) = -1. \end{aligned}$$

Proof. Assume that $u, v \in N$ are such that $p_{uv} > 0$ (we know by (36) that $p_{uv} \geq 0$). Then for any subset $S \subseteq N$ such that $u, v \in S$, we have $g(S,0) = p(S) > 0$ and hence it must be the case that $g(S,1) = 0$ if $S \neq N$ and $g(N,1) = -1$.

This means that the restriction of the function $g(x,1)$ to the subcube $\{x : x_u = x_v = 1\}$, say $d(x)$, is equal to the negative monomial M_{n-2}. But when $n \geq 5$, $d(x)$ is quadratic whereas the degree of M_{n-2} is at least three, a contradiction. When $n = 4$, say $N = \{u, v, t, w\}$, the unicity of the polynomial representation of $M_{n-2} = M_2$ implies that $d(x) = -x_t x_w$, i.e., $p_{tw} = -1$ which contradicts (36).

Thus, we have $p_{uv} = 0$ for all $u, v \in N$. Finally, the claimed inequalities and equality follow from Equations (28)–(29). $\square$

The previous relations allow us to establish a first case of Theorem 8.

Claim 3 *The statement of Theorem 8 holds when $N^+ = \emptyset$ and $N = N^0$.*

Proof. If $N = N^0$, then $c(S) = 0$ for all $S \subseteq N$ by definition and, by Claim 2, $p(S) = 0$ for all $S \subseteq N$, and $a = -1 - b(N)$. Therefore, we can write

$$g(x, y) \;=\; (-1 - b(N))y + \sum_{u \in N} b_u x_u y.$$

Since $s_n(x, y) = (n-1)y - \sum_{u \in N} x_u y$, we obtain

$$g(x, y) - s_n(x, y) \;=\; (-n - b(N))y + \sum_{u \in N} (b_u + 1) x_u y \;=\; \sum_{u \in N} (-1 - b_u) y \overline{x}_u. \quad (38)$$

The relations $a + b(N \setminus \{u\}) \geq 0$ and $a + b(N) = -1$ imply that $b_u \leq -1$ for all $u \in N$. Hence, the right-hand side of (38) is always nonnegative, and if g is prime, then it must be the case that $g = s_n$. $\square$

From now on, let us assume that $|N^+| \geq 1$. Consider $u \in N^0$ and $i \in N^+$. We get $g(\{u, i\}, 0) = c_i + p_{ui}$, and in light of (34), $g(\{u, i\}, 1) = b_u + p_{ui}$. Thus, we can write $N^0 \times N^+ = E_B \cup E_C$, where

$$E_B = \{(u, i) \;\mid\; u \in N^0,\; i \in N^+, p_{ui} = -b_u\}, \text{ and} \quad (39)$$

$$E_C = \{(u, i) \;\mid\; u \in N^0,\; i \in N^+, p_{ui} = -c_i\}. \quad (40)$$

We show next some properties of E_B, E_C, which will be useful to complete the proof of the main theorem.

We use several times the following identity: when $u \in N^0$ and $i, j \in N^+$, since $p_{ij} = a$ by (37), we have

$$g(\{u, i, j\}, y) = (a + b_u + b_i + b_j)y + c_i + c_j + p_{ui} + p_{uj} + a. \tag{41}$$

Claim 4 *For all $u \in N^0$, we have either $\{u\} \times N^+ \subseteq E_B$, or $\{u\} \times N^+ \subseteq E_C$.*

Proof. Assume that this is not the case, so that there exist $u \in N^0$ and $i, j \in N^+$ such that $(u, i) \in E_B$ and $(u, j) \in E_C$. Then, since $|N| > 3$, we have $0 \leq g(\{u, i, j\}, 1)$. By (34) we have $a + b_i + c_i = a + b_j + c_j = 0$, by (40) $c_j + p_{uj} = 0$, and by (39) $b_u + p_{ui} = 0$. Thus, (41) yields $0 \leq g(\{u, i, j\}, 1) = a + b_j = -c_j$. But this contradicts $j \in N^+$. □

Consider the sets

$$B = \{u \in N^0 \mid \{u\} \times N^+ \subseteq E_B\}, \text{ and} \tag{42}$$
$$C = \{u \in N^0 \mid \{u\} \times N^+ \subseteq E_C\}. \tag{43}$$

The proof of Claim 4 actually establishes the following statement.

Claim 5 *$B \cup C = N^0$ and, if $|N^+| \geq 2$, then $B \cap C = \emptyset$.*

Thus, (B, C) forms a partition of N^0 when $|N^+| \geq 2$. But this is not necessarily true when $|N^+| = 1$. Let us now establish some auxiliary properties of the sets B and C.

Claim 6 *If $|N^+| \geq 1$, then $|C| \leq 1$, and either $B \cap C = \emptyset$ or $B = N^0$.*

Proof. Assume that $i \in N^+$ and $u, v \in C, u \neq v$. Then $b_u \geq c_i = -p_{ui}$ and $b_v \geq c_i = -p_{vi}$. Hence, $a + b_u + b_v + p_{uv} \geq a + 2c_i + p_{uv} > p_{uv}$ and by (36), we must have $p_{uv} = 0$. Then, from (28), $0 \leq g(\{u, v, i\}, 0) = c_i + p_{uv} + p_{ui} + p_{vi} = -c_i$, which contradicts the definition of N^+. This proves that $|C| \leq 1$.

If $B \cap C \neq \emptyset$, then $C \subseteq B$, and hence $B = N^0$. □

Claim 7 *If $|N^+| \geq 1$, $u \in B$, $v \in C \setminus B$, and $u \neq v$, then $p_{uv} = b_u$.*

Proof. Let $i \in N^+$. According to the definitions, $g(\{u, v, i\}, y) = (a + b_u + b_v + b_i)y - b_u + p_{uv}$. By definition of C, $b_v - c_i = b_v + p_{vi}$ and since $v \notin B$, $b_v + p_{vi} > 0$.

By (34) we have $a + b_i = -c_i$, and hence we get $g(\{u, v, i\}, 1) = b_v - c_i + p_{uv}$. Thus, $g(\{u, v, i\}, 1) > 0$, since $p_{uv} \geq 0$ by (36). Consequently, $g(\{u, v, i\}, 0) = -b_u + p_{uv} = 0$. This proves the claim. □

Claim 8 *If $|N^+| \geq 2$, then $B = \emptyset$ and $C = N^0$.*

Proof. Assume by contradiction that $B \neq \emptyset$. Let us consider an arbitrary $u \in B$ and $i, j \in N^+$, $i \neq j$. Then, we have $g(\{u, i, j\}, 1) = -b_u$ by (41), (34), (37) and the definition of B. Thus we have $-b_u \geq 0$ from which $g(\{u, i, j\}, 0) = c_i + c_j - 2b_u + a > 0$ follows, implying that we must have $g(\{u, i, j\}, 1) = -b_u = 0$. Hence $p_{ui} = 0$ for all $i \in N^+$, by definition of B. Also, for all $v \in C$, Claim 5 implies that $v \notin B$, and by Claim 7, $p_{uv} = b_u = 0$.

Assume now that $|B| = 1$, $B = \{u\}$. Then, all terms of (27) containing x_u vanish, since $b_u = c_u = p_{ui} = p_{uv} = 0$ for all $i \in N^+$, $v \in C$. Thus, x_u does not appear in g, a contradiction with the fact that M_n depends on all its variables.

On the other hand, if $|B| > 1$ and $v \in B$, $v \neq u$, then $g(\{u, v, i, j\}, y) = (a + b_i + b_j)y + c_i + c_j + a + p_{uv}$. Here a and p_{uv} are nonnegative by (30) and (36), and c_i and c_j are both positive by the definition of N^+, therefore $g(\{u, v, i, j\}, 0) > 0$. Thus, $g(\{u, v, i, j\}, 1) = p_{uv} \leq 0$ follows by (28) and (29). Since $p_{uv} \geq 0$ by (36), $p_{uv} = 0$ follows. Consequently, $b_u = p_{uv} = 0$ follows for all $u \in N^0$ and $v \neq u$, implying again that x_u does not play any role in g, which is a contradiction and proves our claim. □

Claim 9 *If $|N^+| \geq 2$, then $|C| = 1$, $|N^+| = n - 1$, and $a = b_i + c_i = p_{ij} = 0$ for all $i, j \in N^+$.*

Proof. When $|N^+| \geq 2$, Claim 6 and Claim 8 together imply that $B = \emptyset$, $C = N^0$, and $|C| \leq 1$. Since $N^0 \neq \emptyset$ by Claim 1, it follows that $|C| = 1$ and $|N^+| = n - 1$.

We assumed $|N| \geq 4$. So, let $i, j, k \in N^+$ be three distinct indices. Then $g(\{i,j,k\}, 0) = c_i + c_j + c_k + 3a > 0$ by (37), by definition of N^+ and by (30). Thus, we must have $g(\{i,j,k\}, 1) = 0$ by (28). By (34), this implies $a = 0$, and the claim follows by (37). □

Claim 10 *If $g(x, y)$ is a quadratization of M_n with $|N^+| \geq 2$, then $h(x, y) = g(x, \overline{y})$ is another quadratization of M_n with either $|N^+| = 1$ and $|B| = n-1$, or $N^+ = \emptyset$ and $N = N^0$.*

Proof. This follows from the definitions and from Claim 9. □

In view of Claim 3 and Claim 9, up to switching the y-variable, we are left with the case $|N^+| = 1$.

Claim 11 *If $N^+ = \{i\}$, then $p_{uv} = 0$ for all $u, v \in B$.*

Proof. Let us assume there exist $u, v \in B$ such that $p_{uv} > 0$ (we know by (36) that $p_{uv} \geq 0$.) Then $g(\{u,v,i\}, 1) = (a + b_u + b_v + b_i) + c_i + p_{uv} - b_u - b_v = p_{uv} > 0$ by (34) and by the definition of B. Thus, $g(\{u,v,i\}, 0) = c_i + p_{uv} - b_u - b_v = 0$ follows by (28). On the other hand, we have $g(\{u,v\}, 0) = p_{uv} > 0$ and thus $g(\{u,v\}, 1) = a + b_u + b_v + p_{uv} = 0$ follows again by (28). Adding these two equalities, we get $a + c_i + 2p_{uv} = 0$ which is impossible since $a \geq 0$, $c_i > 0$ and $p_{uv} > 0$. □

Claim 12 *If $N^+ = \{i\}$, then $|B| = n - 1$. Furthermore, we have*

$$c_i = b(B) - 1, \text{ and} \tag{44}$$

$$c_i \geq b(S) \text{ for all subsets } S \subseteq B, S \neq B. \tag{45}$$

Proof. Assume first that $|B| < n-1$. It follows from Claim 6 that $|C| = 1$ and $B \cap C = \emptyset$. Let $C = \{w\}$. We obtain

$$g(N, y) = (a + b(B) + b_w + b_i)y + c_i + p(B) + \sum_{u \in B} p_{uw} + \sum_{u \in B} p_{ui} + p_{wi}. \quad (46)$$

Now, $a + b_i = -c_i$ by (34), $p(B) = 0$ by Claim 11, $\sum_{u \in B} p_{uw} = \sum_{u \in B} b_u$ by Claim 7, $\sum_{u \in B} p_{ui} = -\sum_{u \in B} b_u$ by definition of B, and $p_{wi} = -c_i$ by definition of C. Hence,

$$g(N, y) = (b(B) + b_w - c_i)y. \quad (47)$$

In view of Claim 7 and of (36), $b_u = p_{uw} \geq 0$ for all $u \in B$. Moreover, $g(\{w, i\}, 1) = b_w + p_{wi} = b_w - c_i$ by definition of C, and hence $b_w - c_i \geq 0$. This implies that $g(N, y) \geq 0$ for all y, contradicting (29).

Thus, $|B| = n-1$. In this case we obtain $g(N, 1) = 0$ by definition of B, and thus we must have $g(N, 0) = c_i - b(B) = -1$. Furthermore, for any subset $S \subseteq B$, $S \neq B$ we have $g(S, 0) = c_i - b(S) \geq 0$. □

We are now ready to prove the remaining case of Theorem 8.

Claim 13 *The statement of Theorem 8 holds when $|N^+| = 1$.*

Proof. In view of Claim 12, we can assume that $N^+ = \{n\}$ and that $B = \{1, 2, ..., n-1\}$. By (34), by the definition of B and by Claim 11, we have $b_n = -a - c_n$, $p_{nu} = -b_u$ for all $u \in B$, and $p_{uv} = 0$ for all $u, v \in B$. Thus,

$$g(x, y) \;=\; ay\overline{x}_n + c_n x_n \overline{y} + \sum_{u \in B} b_u x_u (y - x_n).$$

Since $s_n^+(x, \overline{y}) = (n-2)x_n\overline{y} + \sum_{u \in B} x_u(y - x_n)$, we get

$$g(x, y) - s_n^+(x, \overline{y}) = ay\overline{x}_n + (c_n - n + 2)x_n\overline{y} + \sum_{u \in B} (b_u - 1)x_u(y - x_n).$$

By (44), we have $\sum_{u\in B}(b_u - 1) = c_n - n + 2$. Hence, we can write

$$\begin{aligned} g(x,y) - s_n^+(x,\overline{y}) &= ay\overline{x}_n + \sum_{u\in B}(b_u-1)[x_u(y-x_n) + x_n\overline{y}] \\ &= ay\overline{x}_n + \sum_{u\in B}(b_u-1)[yx_u\overline{x}_n + \overline{y}\,\overline{x}_u x_n]. \end{aligned}$$

The relations (44)–(45) imply that $b_u \geq 1$ for all $u \in B$. Hence, $g(x,y) - s_n^+(x,\overline{y})$ is always nonnegative, and this completes the proof of the theorem. □

6 Conclusions

This paper initiates a systematic study of quadratizations of pseudo-Boolean functions. Matching lower and upper bounds are established for the number of auxiliary variables required in smallest possible quadratizations of arbitrary functions and of of degree-d polynomials. Similar bounds are also derived for the restricted, but frequently-considered class of y-linear quadratizations. Future research should provide a numerical assessment of the quality of the constructive quadratization procedures leading to the upper bounds.

Theorem 8 provides a complete characterization of prime quadratizations of negative monomials using only one auxiliary variable. Although this may seem to be a rather modest result, its proof turns out to be quite intricate (perhaps for lack of better ideas). The case of positive monomials appears to be harder, since we do not even know the minimum number of auxiliary variables required in a shortest quadratization of $\prod_{i=1}^n x_i$ (see Anthony et al. [1] and Ishikawa [32] for the best known upper bounds).

Many other questions about quadratizations remain open. In particular, the computational complexity of minimizing the number of auxiliary variables for a given pseudo-Boolean function is unknown. The problem seems to be very high in the polynomial hierarchy. We conjecture that it is Σ_3^p-complete.

acknowledgements

We are grateful to Zoltán Szigeti and György Turán for pointers on 2-bases and on representations of symmetric functions, and to the reviewers for their valuable comments. The second author thanks the National Science Foundation (Grant IIS-1161476) for partial support. The third author was partially funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office (grant P7/36) and by a sabbatical grant from FNRS. The fourth author thanks the joint CAPES (Brazil)/Fulbright (USA) fellowship process BEX-2387050/15061676 and the São Paulo Science Foundation (FAPESP) process 2014/23269-8 for partial suport.

References

[1] M. Anthony, E. Boros, Y. Crama and M. Gruber, Quadratization of symmetric pseudo-Boolean functions, *Discrete Applied Mathematics* 203 (2016) 1–12.

[2] A. Billionnet and S. Elloumi, Using a mixed integer quadratic programming solver for the unconstrained quadratic 0-1 problem, *Mathematical Programming* 109 (2007) 55–68.

[3] E. Boros and A. Gruber, On quadratization of pseudo-Boolean functions, Working paper, 2011.

[4] E. Boros and P.L. Hammer, Pseudo-Boolean optimization, *Discrete Applied Mathematics* 123 (2002) 155–225.

[5] E. Boros, P.L. Hammer, R. Sun and G. Tavares, A max-flow approach to improved lower bounds for quadratic unconstrained binary optimization (QUBO), *Discrete Optimization* 5 (2008) 501–529.

[6] E. Boros, P.L. Hammer and G. Tavares, Local search heuristics for quadratic unconstrained binary optimization, *Journal of Heuristics* 13 (2007) 99–132.

[7] Y. Boykov, O. Veksler and R. Zabih, Fast approximate energy minimization via graph cuts, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23 (2001) 1222–1239.

[8] Ch. Buchheim and G. Rinaldi, Efficient reduction of polynomial zero-one optimization to the quadratic case, *SIAM Journal on Optimization* 18 (2007) 1398–1413.

[9] Ch. Buchheim and G. Rinaldi, Terse integer linear programs for Boolean optimization, *Journal on Satisfiability, Boolean Modeling and Computation* 6 (2009) 121–139.

[10] S. Burer and A.N. Letchford, Non-convex mixed-integer nonlinear programming: A survey, *Surveys in Operations Research and Management Science* 17 (2012) 97–106.

[11] M. Conforti, G. Cornuéjols and G. Zambelli, Extended formulations in combinatorial optimization, *4OR* 8 (2010) 1–48.

[12] Y. Crama and P.L. Hammer, Pseudo-Boolean optimization, in: P.M. Pardalos and M.G.C. Resende, eds., *Handbook of Applied Optimization*, Oxford University Press, 2002, pp. 445–450.

[13] Y. Crama and P.L. Hammer, *Boolean Functions: Theory, Algorithms, and Applications*, Cambridge University Press, New York, N.Y., 2011.

[14] Y. Crama and E. Rodríguez-Heck, Short prime quadratizations of cubic negative monomials. Research report, July 2014. `http://hdl.handle.net/2268/170649`

[15] D. Ellis and B. Sudakov, Generating all subsets of a finite set with disjoint unions, *Journal of Combinatorial Theory, Series A* 118 (2011) 2319–2345.

[16] A. Fix, Reductions for rewriting QPBFs with spanning trees. Unpublished notes, 2011.

[17] A. Fix, A. Gruber, E. Boros and R. Zabih, A graph cut algorithm for higher-order Markov random fields, Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV), pages 1020–1027.

[18] A. Fix, A. Gruber, E. Boros and R. Zabih, A hypergraph-based reduction for higher-order Markov random fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (2015) 1387–1395.

[19] R. Fortet, L'algèbre de Boole et ses applications en recherche opérationnelle, *Cahiers du Centre d'Etudes de Recherche Opérationnelle* 1 (1959) 5–36.

[20] R. Fortet, Applications de l'algèbre de Boole en recherche opérationnelle, *Revue Française de Recherche Opérationnelle* 4 (1960) 17–26.

[21] D. Freedman and P. Drineas, Energy minimization via graph cuts: Settling what is possible, in: *IEEE Conference on Computer Vision and Pattern Recognition* (2) (2005) pp. 939-946.

[22] Y. Frein, B. Lévêque and A. Sebő, Generating all sets with bounded unions, *Combinatorics, Probability and Computing* 17 (2008) 641–660.

[23] Z. Füredi and G.O.H. Katona, 2-Bases of quadruples, *Combinatorics, Probability and Computing* 15 (2006) 131–141.

[24] F. Glover, B. Alidaee, C. Rego and G. Kochenberger, One-pass heuristics for large-scale unconstrained binary quadratic problems, *European Journal of Operational Research* 137 (2002) 272–287.

[25] F. Glover and J.-K. Hao, Efficient evaluations for solving large 0–1 unconstrained quadratic optimisation problems, *International Journal of Metaheuristics* 1 (2010) 3–10.

[26] P.L. Hammer, P. Hansen and B. Simeone, Roof duality, complementation and persistency in quadratic 0–1 optimization, *Mathematical Programming* 28 (1984) 121–155.

[27] P.L. Hammer and S. Rudeanu, *Boolean Methods in Operations Research and Related Areas,* Springer, Berlin, 1968.

[28] P. Hansen, B. Jaumard and V. Mathon, Constrained nonlinear 0-1 programming, *ORSA Journal on Computing* 5 (1993) 97–119.

[29] P. Hansen and Ch. Meyer, Improved compact linearizations for the unconstrained quadratic 01 minimization problem, *Discrete Applied Mathematics* 157 (2009) 1267–1290.

[30] C. Helmberg and F. Rendl, Solving quadratic (0,1)-problems by semidefinite programs and cutting plane, *Mathematical Programming* 82 (1998) 291–315.

[31] R. Hemmecke, M. Köppe, J. Lee and R. Weismantel, Nonlinear integer programming, in M. Jünger, T.M. Liebling, D. Naddef, G.L. Nemhauser, W.R. Pulleyblank, G. Reinelt, G. Rinaldi and L.A. Wolsey, eds., *50 Years of Integer Programming*, Springer-Verlag, Berlin Heidelberg, 2010, pp. 561–618.

[32] H. Ishikawa, Transformation of general binary MRF minimization to the first-order case, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(6) (2011) 1234–1249.

[33] F. Kahl and P. Strandmark, Generalized roof duality, *Discrete Applied Mathematics* 160 (2012) 2419–2434.

[34] J.H. Kappes et al., A comparative study of modern inference techniques for discrete energy minimization problems, in: *IEEE Conference on Computer Vision and Pattern Recognition* CVPR'13 (2013) pp. 1328–1335.

[35] V. Kolmogorov, Generalized roof duality and bisubmodular functions, *Discrete Applied Mathematics* 160 (2012) 416–426.

[36] V. Kolmogorov and C. Rother, Minimizing non-submodular functions with graph cuts - A review, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29 (2007) 1274–1279.

[37] V. Kolmogorov and R. Zabih, What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(2) (2004) 147–159.

[38] K. Maghout, Sur la détermination des nombres de stabilité et du nombre chromatique d'un graphe, *Comptes Rendus de l'Académie des Sciences de Paris* 248 (1959) 3522–3523.

[39] K. Maghout, Applications de l'algèbre de Boole à la théorie des graphes et aux programmes linéaires et quadratiques, *Cahiers du Centre d'Etudes de Recherche Opérationnelle* 5 (1963) 21–99.

[40] P. Merz and B. Freisleben, Greedy and local search heuristics for the unconstrained binary quadratic programming problem, *Journal of Heuristics* 8(2002) 197–213.

[41] S. Ramalingam, Ch. Russell, L. Ladický and Ph.H.S. Torr, Efficient minimization of higher order submodular functions using monotonic Boolean functions, 2011. `ArXiv:1109.2304v1`

[42] I.G. Rosenberg, Reduction of bivalent maximization to the quadratic case, *Cahiers du Centre d'Etudes de Recherche Opérationnelle* 17 (1975), 71–74.

[43] C. Rother, P. Kohli, W. Feng and J. Jia, Minimizing sparse higher order energy functions of discrete variables, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1382–1389.

[44] C. Rother, V. Kolmogorov, V. Lempitsky and M. Szummer, Optimizing binary MRFs via extended roof duality, in: *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[45] H.D. Sherali and W.P. Adams, *A Reformulation-Linearization Technique for Solving Discrete and Continuous Nonconvex Problems,* Kluwer, Dordrecht, 1999.

[46] A. Sidorenko, What we know and what we do not know about Turán numbers, *Graphs and Combinatorics* 11 (1995) 179–199.

[47] S. Živný, D.A. Cohen and P.G. Jeavons, The expressive power of binary submodular functions, *Discrete Applied Mathematics* 157 (2009) 3347–3358.

[48] S. Živný and P. G. Jeavons, Classes of submodular constraints expressible by graph cuts, *Constraints* 15 (2010) 430–452.