Journal on Satisfiability, Boolean Modeling and Computation 6 (2008) 1-12

Matched Formulas and Backdoor Sets*

Stefan Szeider stefan.szeider@durham.ac.uk
Department of Computer Science
Durham University
Durham DH1 3LE, England, United Kingdom

Abstract

We demonstrate hardness results for the detection of small backdoor sets with respect to base classes $\mathcal{M}_r$ of CNF formulas with maximum deficiency $\leq r$ ($\mathcal{M}_0$ is the class of matched formulas). One of the results applies also to a wide range of base classes with added 'empty clause detection' as recently considered by Dilkina, Gomes, and Sabharwal. We obtain the hardness results in the framework of parameterized complexity, considering the upper bound on the size of smallest backdoor sets as the parameter. Furthermore we compare the generality of the parameters maximum deficiency and the size of a smallest $\mathcal{M}_r$-backdoor set.

KEYWORDS: *maximum deficiency, backdoor sets, parameterized complexity*

Submitted December 2007; revised June 2008; published September 2008

1 Introduction and Background

1.1 Matched Formulas

A CNF formula is *matched* if one can match each clause to a 'private' variable that occurs in the clause such that different clauses are matched to different variables. Matched CNF formulas are satisfiable since one can satisfy each clause independently by choosing the right truth value for its private variable. Moreover, such formulas can be recognized efficiently by bipartite matching algorithms. Matched formulas play a prominent role in several theoretical investigations. For example, they were used in Tovey's classical paper on 3SAT with bounded occurrence of variables [23], and in Tarsi's Lemma on the clause-variable difference of minimal unsatisfiable formulas [1]. In a certain sense, matched formulas are more numerous then formulas belonging to other well-known tractable classes such as Horn and renamable Horn formulas [10]. The classes of biclique-satisfiable and var-satisfiable formulas properly contain all matched formulas, but the recognition problems for these two classes are intractable [22].

The notion of *maximum deficiency*, first used by Franco and Van Gelder [10] in the context of CNF formulas, allows to extend the nice properties of matched CNF formulas to more general classes of formulas. The maximum deficiency of a CNF formula F, denoted by $md(F)$, is the number of clauses remaining without a private variable in an optimal matching (more precisely, in a maximum cardinality matching in the incidence graph associated with

∗ A preliminary and shortened version appeared in the Proceedings of SAT 2007.

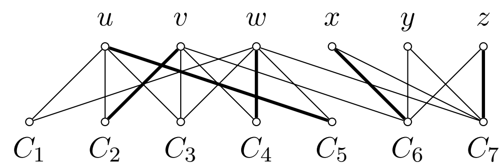

Figure 1. Incidence graph associated with the CNF formula $F = \{C_1, \dots, C_7\}$, $C_1 = \{u, w\}$, $C_2 = \{\neg u, v\}$, $C_3 = \{u, \neg v, \neg w\}$, $C_4 = \{v, \neg w\}$, $C_5 = \{\neg u, w\}$, $C_6 = \{\neg v, x, y, z\}$, $C_7 = \{\neg x, \neg y, \neg z\}$ (each variable is adjacent to all clauses it occurs in). Bold edges indicate a maximum cardinality matching. This matching assigns to C_2, C_4, C_5, C_6, C_7 the private variables v, w, u, x, z, respectively. Clauses C_1 and C_3 have no private variable according to this matching, therefore the maximum deficiency of F is $7 - 5 = 2$.

the CNF formula, see Figure 1 for an example). The term 'maximum deficiency' is motivated by the equality $md(F) = \max_{F' \subseteq F} d(F')$ which follows from Hall's Theorem; here $d(F')$ denotes the *deficiency* of F', the difference between the number of clauses and the number of variables of F'. Let us denote the class of CNF formulas with maximum deficiency at most r by $\mathcal{M}_r$ (thus $\mathcal{M}_0$ denotes the class of matched formulas).

Matchings can be used to simplify CNF formulas and to compute a normal form. For example, from the formula F in Figure 1 one can remove the clauses C_6 and C_7 since the indicated matching yields the assignment that sets x to 1 and z to 0, and these two variables do not occur in the remaining clauses (the assignment is 'autark' [15]). This leaves us with the CNF formula $F' = \{C_1, \dots, C_5\}$ where F and F' are satisfiability-equivalent and $d(F') = md(F') = md(F)$. In general, for any given CNF formula F, one can find in polynomial time a subset F' such that F and F' are satisfiability-equivalent and $d(F') = md(F') = md(F)$; if $F = F'$ then F is called 'matching lean.' Kullmann [15] provides an in-depth study of autark assignments and the corresponding notion of lean CNF formulas. For our constructions below it is convenient to consider matching lean formulas because their deficiency is easy to compute.

A CNF formula is *minimal unsatisfiable* if it is unsatisfiable and each proper subset is satisfiable. Clearly minimal unsatisfiable formulas are matching lean. Kleine Büning [13] initiates the study of minimal unsatisfiable formulas parameterized by their deficiency; this study is also extended to quantified and non-Boolean formulas [14, 16]. Fleischner, Kullmann, and Szeider [8] show that for every constant r, one can decide the satisfiability of CNF formulas with maximum deficiency r in time $O(L \cdot n^{r+1/2})$; here L denotes the length and n the number of variables of the given CNF formula. As a consequence, minimal unsatisfiable formulas with deficiency bounded by a constant r can be recognized with the same polynomial bound $O(L \cdot n^{r+1/2})$. The order of the polynomial time bound depends on r which makes the algorithm infeasible for larger inputs, even if r is small (say $r = 5$). Szeider [21] develops an algorithm that overcomes this limitation: the algorithm decides satisfiability of CNF formulas with n variables and maximum deficiency r in time $O(2^r n^3)$ and recognizes minimal unsatisfiable formulas with deficiency r in time $O(2^r n^4)$. Thus, the satisfiability problem is *fixed-parameter tractable* with respect to the parameter maximum deficiency, since the degree of the polynomial is now independent of the parameter r. We discuss some basic concepts of parameterized complexity in Section 2.1; for more background information we refer the reader to other sources [6, 9, 17].

1.2 Backdoor Sets

Williams, Gomes, and Selman [24] introduce the notion of backdoor sets for analyzing the 'heavy-tailed' behavior of backtrack search algorithms. Basically, a backdoor set is a (small) set of variables of a given CNF formula that, when instantiated, puts the formula into a tractable base class. In this paper we consider backdoor sets with respect to the base classes $\mathcal{M}_r$ as defined above. The size of such backdoor sets together with the base class level r provide a 'two-layered' parameterization of the satisfiability problem. Bidyuk and Dechter [2] consider a somewhat similar two-layered parameterization of Bayesian reasoning problems.

Let F be a CNF formula, B a subset of the set of variables of F, and let $\mathcal{C}$ be a base class of tractable instances. If for every truth assignment $\tau : B \to \{0,1\}$, the restriction $F[\tau]$ belongs to $\mathcal{C}$, then B is a *strong $\mathcal{C}$-backdoor set*; if for at least one $\tau : B \to \{0,1\}$ the formula $F[\tau]$ is both satisfiable and belongs to $\mathcal{C}$, then B is a *weak $\mathcal{C}$-backdoor set*. See Section 2.2 for notational conventions on CNF formulas and truth assignments. A variant of strong backdoor sets are *deletion* backdoor sets: B is a deletion backdoor set if the formula $F - B$ belongs to $\mathcal{C}$; $F - B$ denotes the formula obtained from F by removing all literals $x, \neg x$ with $x \in B$ from the clauses of F. If the base class $\mathcal{C}$ is *clause-induced* (that is, if $F \in \mathcal{C}$ implies $F' \in \mathcal{C}$ for all $F' \subseteq F$), then every deletion $\mathcal{C}$-backdoor set is a strong $\mathcal{C}$-backdoor set [19]. Deletion backdoor sets are useful, as for certain base classes it is easier to search for deletion backdoor sets than for strong backdoor sets [19]. To exclude degenerate cases, we consider weak backdoor sets only for base classes $\mathcal{C}$ with $\emptyset \in \mathcal{C}$, and we consider strong and deletion backdoor sets only for base classes $\mathcal{C}$ with $\emptyset, \{\emptyset\} \in \mathcal{C}$.

Considering a base class $\mathcal{C}$ and a CNF formula F, we denote by

$$wb_{\mathcal{C}}(F),\ sb_{\mathcal{C}}(F), \text{ and } db_{\mathcal{C}}(F)$$

the size of a smallest weak $\mathcal{C}$-backdoor set, the size of a smallest strong $\mathcal{C}$-backdoor set, and the size of a smallest deletion $\mathcal{C}$-backdoor set of F, respectively. Note that $wb_{\mathcal{C}}(F)$ is only defined if F is satisfiable. We always have $sb_{\mathcal{C}}(F) \leq db_{\mathcal{C}}(F)$ if $\mathcal{C}$ is clause-induced.

A base class $\mathcal{C}$ is *self-reducible* if $F \in \mathcal{C}$ implies $F[x=0], F[x=1] \in \mathcal{C}$ (see [5]). The base classes $\mathcal{M}_r$ are not self-reducible in this sense, but in a slightly wider sense: $F \in \mathcal{M}_r$ implies $F[x=0], F[x=1] \in \mathcal{M}_{r+1}$, and if F is matching lean, then $F \in \mathcal{M}_r$ implies $F[x=0], F[x=1] \in \mathcal{M}_r$ (see [15]). As mentioned above, we can always find efficiently a matching-lean subset F' of a given CNF formula F such that F and F' are satisfiability-equivalent.

If we know a strong backdoor set B of a CNF formula F, then we can decide the satisfiability of F by checking the satisfiability of at most $2^{|B|}$ formulas $F[\tau]$ that belong to the base class (in some cases it suffices to check significantly fewer than $2^{|B|}$ formulas [20]). Thus it is interesting to find for a given formula a small backdoor set, say one of size bounded by some fixed integer k. Of course we can consider all sets of variables up to size k, and check whether at least one of them is a backdoor set. This exhaustive search requires time of order n^k for CNF formulas with n variables, thus it becomes infeasible for large n even if k is reasonably small. Whether we can do significantly better than exhaustive search can be studied within the framework of parameterized complexity (see Section 2.2). Nishimura, Ragde, and Szeider [18] show that, with respect to the base classes Horn and 2CNF, strong and deletion backdoor sets coincide. Furthermore, they show that for parameter k, deciding whether $wb_{\text{Horn}}(F) \leq k$ and deciding whether $wb_{2\text{CNF}}(F) \leq k$ are not fixed-parameter tractable problems (under reasonable complexity theoretic assumptions) and that deciding whether $sb_{\text{Horn}}(F) \leq k$ and $sb_{2\text{CNF}}(F) \leq k$ are fixed-parameter tractable problems. However, if k is not a parameter but just a part of the input, then the problems are NP-complete [3, 18].

Dilkina, Gomes, and Sabharwal [5] suggest to strengthen the concept of strong backdoor sets by means of *empty clause detection*. Let $\mathcal{E}$ denote the class of all CNF formulas that contain the empty clause. For a base class $\mathcal{C}$ we put $\mathcal{C}^{\{\}} = \mathcal{C} \cup \mathcal{E}$; we call $\mathcal{C}^{\{\}}$ the base class obtained from $\mathcal{C}$ by adding empty clause detection. Formulas often have much smaller strong $\mathcal{C}^{\{\}}$-backdoor sets than strong $\mathcal{C}$-backdoor sets [5]. Dilkina et al. show that, given a CNF formula F and an integer k, determining whether F has a strong $\text{Horn}^{\{\}}$-backdoor set of size k, is both NP-hard and co-NP-hard (here k is considered just as part of the input and not as a parameter). Thus, the non-parameterized search problem for strong Horn-backdoor sets gets harder when empty clause detection is added. We demonstrate that for many base classes, including $\mathcal{M}_r$, Horn, and 2CNF, the parameterized search problem also gets harder when empty clause detection is added.

1.3 Results

We compare the parameters md, $wb_{\mathcal{M}_r}$, $sb_{\mathcal{M}_r}$, and $db_{\mathcal{M}_r}$ with each other, and we determine the complexity of recognizing CNF formulas with small $wb_{\mathcal{M}_r}$, $sb_{\mathcal{M}_r}$, and $db_{\mathcal{M}_r}$, respectively. We obtain the following results.

1. Let $r \geq 0$. For every satisfiable CNF formula F we have $wb_{\mathcal{M}_r}(F) \leq md(F)$. The difference between $wb_{\mathcal{M}_r}(F)$ and $md(F)$ can be arbitrarily large. (Therefore we say that $wb_{\mathcal{M}_r}$ is a 'more general parameter' than md.)

2. Let $r > 0$. There are CNF formulas with arbitrarily large $sb_{\mathcal{M}_r}$ and constant md. On the other hand, there are CNF formulas with arbitrarily large $md(F)$ and constant $sb_{\mathcal{M}_r}$. (Thus we say the parameters md and $sb_{\mathcal{M}_r}$ are 'incomparable.') The same holds true when considering $db_{\mathcal{M}_r}$ instead of $sb_{\mathcal{M}_r}$.

3. Deciding whether $wb_{\mathcal{M}_r}(F) \leq k$, $sb_{\mathcal{M}_r}(F) \leq k$, or $db_{\mathcal{M}_r}(F) \leq k$, respectively, parameterized by k, is not fixed-parameter tractable (under reasonable complexity-theoretic assumptions).

4. Deciding whether $sb_{\mathcal{M}_r^{\{\}}}(F) \leq k$, parameterized by k, is not fixed-parameter tractable (under reasonable complexity-theoretic assumptions). In fact, this hardness result holds for a wide range of clause-induced base classes with added empty clause detection, and in particular for $\text{Horn}^{\{\}}$ and $2\text{CNF}^{\{\}}$.

2 Preliminaries

2.1 Notation

We consider propositional formulas in conjunctive normal form, *CNF formulas*, represented as sets of clauses. A clause is a set of *literals*, a literal is a variable or a negated variable. For a CNF formula F we denote by $\mathrm{var}(F)$ the set of variables that occur (negated or unnegated) in F. The *length* of F is $\sum_{C\in F}|C|$. A (partial truth) *assignment* is a mapping $\tau : X \rightarrow \{0,1\}$ defined for some set X of variables. Assignments extend in the obvious way to literals by setting $\tau(\neg x) = 1 - \tau(x)$. The *restriction* of F to τ is the CNF formula $F[\tau]$ obtained from F by removing all clauses that contain a literal ℓ with $\tau(\ell) = 1$ and by removing all literals ℓ with $\tau(\ell) = 0$ from the remaining clauses. F is *satisfiable* if $F[\tau] = \emptyset$ for some truth assignment τ. For a set X of variables we write $\overline{X} = \{\, \neg x : x \in X \,\}$. By *deleting* a set X of variables from a CNF formula F we obtain the CNF formula

$$F - X = \{\, C \setminus (X \cup \overline{X}) : C \in F \,\}.$$

2.2 Parameterized Complexity

An instance of a parameterized problem is a pair (I, k) where I is the *main part* and k is the *parameter*; the latter is usually a non-negative integer. A parameterized problem is *fixed-parameter tractable* if instances (I, k) can be solved in time $O(f(k)\|I\|^c)$ where f is a computable function, c is a constant, and $\|I\|$ represents the size of I in a reasonable encoding. FPT denotes the class of all fixed-parameter tractable decision problems.

A *parameterized reduction* (or *fpt-reduction*) is a straightforward extension of a polynomial-time many-one reduction that ensures the parameter for one problem maps into the parameter for the other. More specifically, problem L reduces to problem L' if there is a mapping R from instances of L to instances of L' such that (i) (I, k) is a yes-instance of L if and only if $(I', k') = R(I, k)$ is a yes-instance of L', (ii) $k' = g(k)$ for a computable function g, and (iii) R can be computed in time $O(f(k)\|I\|^c)$ where f is a computable function and c is a constant.

The complexity classes $\mathrm{W}[1] \subseteq \mathrm{W}[2] \subseteq \ldots$ are important for the theory of fixed-parameter intractability. The classes are defined as the closure of certain parameterized problems under parameterized reductions. There is strong theoretical evidence that problems that are hard for classes W[i] are not fixed-parameter tractable. For example FPT = W[1] implies that the Exponential Time Hypothesis fails, that is, the existence of a $2^{o(n)}$ algorithm for n-variable 3SAT [9, 12]. Further evidence for W[i] $\neq$ FPT can be obtained by proof complexity theoretic arguments [4]. We use the framework of parameterized complexity to gain evidence that the problems under consideration are not fixed-parameter tractable. For this purpose it is sufficient to establish W[1]-hardness; whether a problem is actually contained in some class W[i] is less important for our considerations.

In this paper we shall consider the following two problems that are known to be W[1]-complete and W[2]-complete, respectively, [6].

clique

Instance: A graph $G = (V, E)$ and a non-negative integer k.

Parameter: k.

Question: Is there there a set $S \subseteq V$ of k distinct vertices that are pairwise adjacent? (the set S is a *clique* of G).

hitting set

Instance: A collection $\mathcal{S} = \{S_1, \ldots, S_m\}$ of finite sets, a non-negative integer k.

Parameter: k.

Question: Is there a set $H \subseteq \bigcup_{i=1}^{m} S_i$ of size k that hits (intersects with) each S_i, $1 \leq i \leq m$? (such H is a *hitting set* of $\mathcal{S}$).

Each base class $\mathcal{C}$ gives rise to the following parameterized problem.

strong $\mathcal{C}$-backdoor set

Instance: A CNF formula F and a non-negative integer k.

Parameter: k.

Question: Is $sb_{\mathcal{C}}(F) \leq k$?

The problems weak $\mathcal{C}$-backdoor set and deletion $\mathcal{C}$-backdoor set are defined analogously.

3 Maximum Deficiency versus Size of Backdoor Sets

Our first result shows that for satisfiable CNF formulas the size of a smallest weak $\mathcal{M}_r$-backdoor set is a strictly more general parameter than maximum deficiency.

Theorem 1. *Let $r \geq 0$. For every satisfiable CNF formula F we have $wb_{\mathcal{M}_r}(F) \leq wb_{\mathcal{M}_0}(F) \leq md(F)$. There are satisfiable CNF formulas with constant $wb_{\mathcal{M}_r}$ and arbitrarily large md.*

Proof. Since $\mathcal{M}_0 \subseteq \mathcal{M}_r$, every weak $\mathcal{M}_0$-backdoor set is also a weak $\mathcal{M}_r$-backdoor set for each $r \geq 0$. Hence, for showing the theorem it suffices to consider the case $r = 0$. Let F be a satisfiable CNF formula with $md(F) = k$. By a result of Fleischner et al. [8] regarding 'matching assignments' (see also Theorem 1 of [21]) it follows that there exists a set $B \subseteq \mathrm{var}(F)$ of size at most k and an assignment $\tau : B \to \{0, 1\}$ such that $F[\tau]$ is a matched formula. Thus $wb_{\mathcal{M}_r}(F) \leq k$.

For the second part of the theorem we chose an arbitrarily large integer $n > 0$. We take variables $x, y_1, \ldots, y_n$ and consider the CNF formula F consisting of the clauses

$$\{x, y_i\},\ \{\overline{x}, y_i\},\ \text{for } i = 1, \ldots, n.$$

The maximum deficiency of F is $|F| - |\mathrm{var}(F)| = 2n - (n + 1) = n - 1$, however $B = \{x\}$ is evidently a weak $\mathcal{M}_0$-backdoor set of F since $F[x = 1]$ is a matched formula. ☐

Next we show that maximum deficiency and the size of a smallest strong/deletion backdoor set are incomparable parameters. Since all formulas in $\mathcal{M}_0$ are satisfiable, unsatisfiable formulas do not have a strong $\mathcal{M}_0$-backdoor set. Therefore, when considering strong $\mathcal{M}_r$-backdoor sets we assume $r > 0$. Note that already $\mathcal{M}_1$ contains unsatisfiable formulas, for example $\{\emptyset\}$.

Theorem 2. *Let $r > 0$. There are CNF formulas with arbitrarily large $sb_{\mathcal{M}_r}$ that have constant md. On the other hand, there are CNF formulas with arbitrarily large md and constant $sb_{\mathcal{M}_r}$. The same holds true when considering $db_{\mathcal{M}_r}$ instead of $sb_{\mathcal{M}_r}$.*

Proof. Let $r > 0$ be a fixed constant and choose an arbitrarily large integer $n > r$.

We take variables x_i^j for $1 \leq i \leq n$ and $1 \leq j \leq r$. We consider the CNF formula F consisting of the clauses

$$\begin{array}{ll} \{\neg x_1^j, \ldots, \neg x_n^j\} & \text{for } 1 \leq j \leq r \text{ and} \\ \{x_i^j\} & \text{for } 1 \leq i \leq n,\ 1 \leq j \leq r. \end{array}$$

It is easy to see that $md(F) = r$, thus the maximum deficiency of F is constant and independent of the choice of n. Let $B \subseteq \mathrm{var}(F)$ be an arbitrarily chosen set of variables with $0 < |B| < n$. Let $\tau_1 : B \to \{1\}$ be the all-1-assignment on B. Now $F[\tau_1]$ consists of exactly r clauses of the form $\{\neg x_1^j, \ldots, \neg x_n^j\} \setminus \{\, \neg x : x \in B \,\}$, $nr - |B|$ clauses of the form $\{x_i^j\}$ for $x_i^j \notin B$, and the empty clause $\emptyset$. Thus, $md(F[\tau_1]) \geq d(F[\tau_1]) = (r+nr-|B|+1)-(nr-|B|) = r + 1$, and so $F[\tau_1] \notin \mathcal{M}_r$. Consequently, $sb_{\mathcal{M}_r}(F) \geq n$. Since $\mathcal{M}_r$ is clause-induced, $db_{\mathcal{M}_r}(F) \geq sb_{\mathcal{M}_r}(F)$, thus $db_{\mathcal{M}_r}(F) \geq n$ holds as well.

Conversely, the CNF formulas considered in the second part of the proof of Theorem 1 have arbitrarily large *md* but constant $sb_{\mathcal{M}_r}$ and constant $db_{\mathcal{M}_r}$ ($B = \{x\}$ is both a strong and a deletion $\mathcal{M}_0$-backdoor set). □

4 Finding Small Backdoor Sets

If r and k are fixed constants, then we can detect strong/weak/deletion $\mathcal{M}_r$-backdoor sets of size at most k in polynomial time, since we can search through all sets of variables of size at most k (there are $O(n^k)$ possibilities if the total number of variables is n) and check the respective conditions. Next we show that we cannot improve significantly upon exhaustive search, subject to the complexity theoretic assumption FPT $\neq$ W[2].

Theorem 3. *Let r be a fixed non-negative integer. The problems* WEAK, STRONG, *and* DELETION $\mathcal{M}_r$-BACKDOOR SET *are* W[2]*-hard.*

Proof. In this proof we will allow the case $k = 0$ for deletion and strong $\mathcal{M}_r$-backdoor sets so that we can treat all three types of backdoor sets uniformly. We give a parameterized reduction from the W[2]-complete problem HITTING SET (see Section 2.2). Let $(\mathcal{S}, k)$ be an instance of this problem with $\mathcal{S} = \{S_1, \ldots, S_m\}$ and $S_i = \{v_i^1, \ldots, v_i^{q_i}\}$. Let $1 \leq q_i = |S_i|$ for $1 \leq i \leq m$ and $V = \bigcup_{i=1}^m S_i$. Since r is constant we may assume that $n = |V| > r$.

We are going to construct a CNF formula F such that $\mathcal{S}$ has a hitting set of size at most k if and only if F has strong/weak/deletion $\mathcal{M}_r$-backdoor set of size at most k. Our general strategy is to construct for each set S_i a CNF formula F_i' with $md(F_i') = r + 1$,

$S_i \subseteq \mathrm{var}(F_i')$, and to consider the union F of all the formulas F_i'. Every $\mathcal{M}_r$-backdoor set of F must involve a variable of F_i' since $md(F_i') = r+1$. On the other hand, F_i' will be constructed in such a way that deleting any variable of $S_i \subseteq \mathrm{var}(F_i')$ reduces the maximum deficiency of F_i' to 0.

For our construction we use three types of variables: 'lump variables', 'hitting variables,' and 'matching variables.' The *hitting variables* are the elements of V, for the other two types we introduce new variables. For each $i \in \{1,\dots,m\}$ we proceed as follows. Let s denote the smallest integer such that $2^s \geq 2(n+1)$. We take new *lump variables* $y_i^1,\dots,y_i^s$ and form a set F_i of $2(n+1)$ different clauses over these variables. For $j = 0,\dots,q_i$ we define inductively CNF formulas F_i^j as follows:

$$\begin{aligned} F_i^0 &:= F_i, \text{ and} \\ F_i^j &:= \{\, C \cup \{v_i^j\} : C \in F_i^{j-1} \,\} \cup \{\, C \cup \{\neg v_i^1,\dots,\neg v_i^j\} : C \in F_i \,\} \text{ for } j > 0. \end{aligned}$$

Clearly $md(F_i^{q_i}) \geq d(F_i^{q_i}) > r$. We take a set $M_i = \{z_i^1,\dots,z_i^{d_i}\}$ of $d_i = md(F_i^{q_i}) - r - 1$ new *matching variables* and put

$$F_i' = \{\, C \cup M_i : C \in F_i^{q_i} \,\}.$$

Each variable of F_i' occurs in each clause of F_i'. Hence it is easy to see that the maximum deficiency of F_i' is exactly $r+1$.

Finally, we obtain the CNF formula $F = \bigcup_{i=1}^m F_i'$. Observe that for $i \neq j$, F_i and F_j do possibly share hitting variables, but do not share any lump variables or matching variables. Clearly F can be constructed from $\mathcal{S}$ in polynomial time.

We show that the following statements are equivalent:

1. $\mathcal{S}$ has a hitting set of size at most k.

2. F has a weak $\mathcal{M}_r$-backdoor set of size at most k.

3. F has a strong $\mathcal{M}_r$-backdoor set of size at most k.

4. F has a deletion $\mathcal{M}_r$-backdoor set of size at most k.

Suppose that $H \subseteq V$ is a hitting set of $\mathcal{S}$. We claim that H is a deletion, strong, and weak $\mathcal{M}_r$-backdoor set of F. Actually, it suffices to show that H is deletion $\mathcal{M}_r$-backdoor set: since $\mathcal{M}_r$ is clause-induced, every deletion backdoor set is also a strong one, and since F is satisfiable (say, by setting all matching variables true) every strong backdoor set is also a weak one. For $i = 1,\dots,m$ let $h_i = |\mathrm{var}(F_i') \cap H|$. Clearly $|F_i' - H| = |F_i'| - 2(n+1)h_i$ and $|\mathrm{var}(F_i' - H)| = |\mathrm{var}(F_i')| - h_i$. From $h_i \geq 1$, $n > r$, and $d(F_i') = r+1$, it follows that $F_i' - H$ has at least n more variables than clauses. Since every variable of $F_i' - H$ occurs in every clause of $F_i' - H$, it follows that $F_i' - H$ is a matched formula. In particular, since at most n of the variables of $F_i' - H$ are hitting variables, we can find for each clause of $F_i' - H$ a private variable that is either a lump or a matching variable. For any pair $1 \leq i \neq j \leq m$ the parts F_i' and F_j' do not share any lump or matching variables. Hence we can combine matchings of the parts $F_i' - H$, $1 \leq i \leq m$, to a matching of the entire formula $F - H$, thus $F - H$ is indeed a matched formula. We conclude that $F - H \in \mathcal{M}_0 \subseteq \mathcal{M}_r$, and so H is a deletion $\mathcal{M}_r$-backdoor set of F. Thus we have shown that statement 1 implies statements 2, 3, and 4.

Now let B be a set of variables of F and let $\tau : B \to \{0, 1\}$ be a truth assignment. Let $F^* \in \{F[\tau], F - B\}$, and assume that $F^* \in \mathcal{M}_r$. We show that $\mathrm{var}(F'_i) \cap B \neq \emptyset$ holds for all $1 \leq i \leq m$. Suppose to the contrary that there is some $1 \leq i \leq m$ such that no variable of F'_i belongs to B. Consequently, F'_i must be a subset of F^*. Now $md(F^*) \geq md(F'_i) = r + 1$ follows, a contradiction to the assumption $F^* \in \mathcal{M}_r$. Hence for each $i \in \{1, \ldots, m\}$ we can pick a variable $x_i \in \mathrm{var}(F'_i) \cap B$, possibly $x_i = x_j$ for $i \neq j$. We define a set $H = \{y_1, \ldots, y_n\}$ of hitting variables as follows. If x_i is a hitting variable, then we put $y_i = x_i$; otherwise we pick $y_i \in S_i$ arbitrarily. It follows now that H is a hitting set of $\mathcal{S}$, and $|H| \leq |B| \leq k$ by construction. Hence each of the statements 2, 3, and 4 implies statement 1. This concludes the proof of Theorem 3. ☐

The reduction in the proof above is actually a polynomial-time many-to-one reduction and does not use the full power of parameterized reductions where it is only required that the reduction is fixed-parameter tractable. Since the non-parameterized version of HITTING SET (where k is part of the input but is not considered as a parameter) is NP-complete [11], it follows that the non-parameterized versions of the problems mentioned in Theorem 3 are NP-hard (again, considering k as part of the input but not as a parameter).

5 Empty Clause Detection

Recall from Section 1.2 that $\mathcal{E}$ denotes the class of all CNF formulas that contain the empty clause, and $\mathcal{C}^{\{\}} = \mathcal{C} \cup \mathcal{E}$ for a base class $\mathcal{C}$. The hardness result we are going to show next holds for all clause-induced base classes $\mathcal{C}$ that are *nontrivial* in the sense that at least one satisfiable CNF formula does not belong to $\mathcal{C}$. Note that the base classes $\mathcal{M}_r$, HORN, and 2CNF are clause-induced and nontrivial.
The following lemma will be useful below.

Lemma 4. *Let F be a CNF formula and $X \subseteq \mathrm{var}(F)$. Then X is a strong $\mathcal{E}$-backdoor set of F if and only if there exists an unsatisfiable subset $F' \subseteq F$ with $\mathrm{var}(F') = X$.*

Proof. Let X be a strong $\mathcal{E}$-backdoor set of F. By definition $\emptyset \in F[\tau]$ holds for all $\tau : X \to \{0, 1\}$, hence $F' = \{ C \in F : \mathrm{var}(C) \subseteq X \}$ is an unsatisfiable subset of F. Conversely, assume $F' \subseteq F$ is unsatisfiable. Then $X = \mathrm{var}(F')$ is a strong $\mathcal{E}$-backdoor set of F. ☐

Theorem 5. *For every nontrivial clause-induced base class $\mathcal{C}$ the problem* STRONG $\mathcal{C}^{\{\}}$-BACKDOOR SET *is* W[1]*-hard.*

Proof. We use a reduction due to Fellows, Szeider, and Wrightson [7] who show that, given a graph G and a positive integer k, one can construct in polynomial time a CNF formula $F(G, k)$ such that for $k' = \binom{k}{2} + 2k$ the following two statements are equivalent.

(1) G contains a clique of size k.

(2) There is an unsatisfiable subset $F' \subseteq F(G, k)$ with $|\mathrm{var}(F')| \leq k'$.

Recall from Section 2.2 that deciding (1) is W[1]-hard (for parameter k); hence deciding (2) is W[1]-hard as well (for parameter k'). Since $\mathcal{C}$ is nontrivial, there exists a satisfiable CNF formula $F \notin \mathcal{C}$. We take several copies $F_1, \dots, F_{k'+1}$ of F such that all the formulas $F(G,k), F_1, \dots, F_{k'+1}$ are mutually variable-disjoint, and we put $F^*(G,k) = F(G,k) \cup \bigcup_{i=1}^{k'+1} F_i$. We show that the following statement is equivalent to statement (2).

(3) $F^*(G,k)$ has a strong $\mathcal{C}^{\{\}}$-backdoor set of size at most k'.

First we show that statement (2) implies statement (3). Assume there is an unsatisfiable subset $F' \subseteq F(G,k)$ with $|\mathrm{var}(F')| \leq k'$. It follows by Lemma 4 that $X = \mathrm{var}(F')$ is a strong $\mathcal{E}$-backdoor set of $F(G,k)$. Clearly X is then a strong $\mathcal{E}$-backdoor set of $F^*(G,k)$ and a strong $\mathcal{C}^{\{\}}$-backdoor set of $F^*(G,k)$.

Next we show that statement (3) implies statement (2). Assume there exists a set $X \subseteq \mathrm{var}(F^*(G,k))$ with $|X| \leq k'$ that is a strong $\mathcal{C}^{\{\}}$-backdoor set of $F^*(G,k)$. Using the pigeonhole principle we conclude that there is some $i \in \{1, \dots, k'+1\}$ such that $X \cap \mathrm{var}(F_i) = \emptyset$. Since $F_i \notin \mathcal{C}$ and since $\mathcal{C}$ is clause-induced, it follows that for every assignment $\tau : X \to \{0,1\}$, the formula $F^*(G,k)[\tau]$ does not belong to $\mathcal{C}$. Thus $\emptyset \in F^*(G,k)[\tau]$ for every $\tau : X \to \{0,1\}$. Hence X is a strong $\mathcal{E}$-backdoor set of $F^*(G,k)$. Next we show that $X' = X \cap \mathrm{var}(F(k,G))$ is a strong $\mathcal{E}$-backdoor set of $F(G,k)$. Let τ^* be a satisfying assignment of $\bigcup_{i=1}^{k'+1} F_i$ (such an assignment exists since all the formulas F_i are satisfiable and mutually variable-disjoint). Consider any assignment $\tau' : X' \to \{0,1\}$. Let $\tau : X \to \{0,1\}$ be the assignment defined by $\tau(x) = \tau'(x)$ for $x \in X'$ and $\tau(x) = \tau^*(x)$ for $x \in X \setminus X'$. Since X is a strong $\mathcal{E}$-backdoor set of $F^*(G,k)$, $\emptyset \in F^*(G,k)[\tau]$ follows. However, since τ does not falsify any clause in $\bigcup_{i=1}^{k'+1} F_i$, $\emptyset \in F(G,k)[\tau] = F(G,k)[\tau']$. Hence X' is a strong $\mathcal{E}$-backdoor set of $F(G,k)$. Consequently statement (2) follows by Lemma 4. □

Corollary 6. *The problems* STRONG $\mathcal{M}_r^{\{\}}$-BACKDOOR SET *($r > 0$),* STRONG HORN$^{\{\}}$-BACKDOOR SET, *and* STRONG 2CNF$^{\{\}}$-BACKDOOR SET *are* W[1]*-hard.*

6 Conclusion

We have considered classes of CNF formulas that properly include the class of matched CNF formulas: CNF formulas with small maximum deficiency and formulas with small $\mathcal{M}_r$-backdoor sets. The results of Section 3 indicate that the second class is more general than the first one when weak backdoor sets are considered, and that the classes are incomparable if strong or deletion backdoor sets are considered. In particular there are classes of formulas with large maximum deficiency and small $\mathcal{M}_r$-backdoor sets. This finding is put into perspective by our hardness results of Section 4 which indicate that it is unlikely that one can find small $\mathcal{M}_r$-backdoor sets efficiently in general. Thus, it is difficult to utilize the power of $\mathcal{M}_r$-backdoor sets algorithmically. We have also shown that the detection of small strong backdoor sets with respect to $\mathcal{M}_r^{\{\}}$ (that is, $\mathcal{M}_r$ with added empty clause detection) is fixed-parameter intractable (Section 5). This intractability result is of independent interest as it holds for a wide range of base classes that includes HORN$^{\{\}}$ and 2CNF$^{\{\}}$. We have focused on a worst-case analysis; it remains open whether SAT solvers would obtain additional power in practice if the existence of backdoor sets with respect to the considered base classes is checked heuristically during the search.

Acknowledgments

The author thanks the reviewers for their suggestions that helped to improve content and presentation of the paper. The research was supported by the EPSRC (EP/E001394/1).

References

[1] Ron Aharoni and Nathan Linial. Minimal non-two-colorable hypergraphs and minimal unsatisfiable formulas. *Journal of Combinatorial Theory, Series A*, **43**:196–204, 1986.

[2] Bozhena Bidyuk and Rina Dechter. Cutset sampling for Bayesian networks. *Journal of Artificial Intelligence Research*, **28**:1–48, 2007.

[3] Yves Crama, Oya Ekin, and Peter L. Hammer. Variable and term removal from Boolean formulae. *Discrete Applied Mathematics*, **75**(3):217–230, 1997.

[4] Stefan Dantchev, Barnaby Martin, and Stefan Szeider. Parameterized proof complexity. In *Proceedings of the 48th Annual Symposium on Foundations of Computer Science (FOCS 2007), October 20-23, 2007, Providence, RI, USA*, pages 150–160. IEEE Press, 2007.

[5] Bistra N. Dilkina, Carla P. Gomes, and Ashish Sabharwal. Tradeoffs in the complexity of backdoor detection. In *Principles and Practice of Constraint Programming - CP 2007, 13th International Conference, CP 2007, Providence, RI, USA, September 23-27, 2007, Proceedings*, **4741** of *Lecture Notes in Computer Science*, pages 256–270. Springer-Verlag, 2007.

[6] Rodney G. Downey and Michael R. Fellows. *Parameterized Complexity*. Monographs in Computer Science. Springer-Verlag, 1999.

[7] Michael R. Fellows, Stefan Szeider, and Graham Wrightson. On finding short resolution refutations and small unsatisfiable subsets. *Theoretical Computer Science*, **351**(3):351–359, 2006.

[8] Herbert Fleischner, Oliver Kullmann, and Stefan Szeider. Polynomial-time recognition of minimal unsatisfiable formulas with fixed clause-variable difference. *Theoretical Computer Science*, **289**(1):503–516, 2002.

[9] Jörg Flum and Martin Grohe. *Parameterized Complexity Theory*, **XIV** of *Texts in Theoretical Computer Science. An EATCS Series*. Springer-Verlag, 2006.

[10] John Franco and Allen Van Gelder. A perspective on certain polynomial time solvable classes of satisfiability. *Discrete Applied Mathematics*, **125**:177–214, 2003.

[11] Michael R. Garey and David R. Johnson. *Computers and Intractability*. W. H. Freeman and Company, New York, 1979.

[12] Russell Impagliazzo, Ramamohan Paturi, and Francis Zane. Which problems have strongly exponential complexity? *Journal of Computer and System Sciences*, **63**(4):512–530, 2001.

[13] Hans Kleine Büning. On subclasses of minimal unsatisfiable formulas. *Discrete Applied Mathematics*, **107**(1–3):83–98, 2000.

[14] Hans Kleine Büning and Xishun Zhao. An extension of deficiency and minimal unsatisfiability of quantified boolean formulas. *Journal on Satisfiability, Boolean Modeling and Computation*, **3**:115–123, 2007.

[15] Oliver Kullmann. Lean clause-sets: Generalizations of minimally unsatisfiable clause-sets. *Discrete Applied Mathematics*, **130**(2):209–249, 2003.

[16] Oliver Kullmann. Constraint satisfaction problems in clausal form: Autarkies and minimal unsatisfiability. Technical Report CSR 8-2007, University of Wales Swansea, 2007.

[17] Rolf Niedermeier. *Invitation to Fixed-Parameter Algorithms*. Oxford Lecture Series in Mathematics and its Applications. Oxford University Press, 2006.

[18] Naomi Nishimura, Prabhakar Ragde, and Stefan Szeider. Detecting backdoor sets with respect to Horn and binary clauses. In *Informal Proc. of SAT 2004 (Seventh International Conference on Theory and Applications of Satisfiability Testing, 10–13 May, 2004, Vancouver, BC, Canada)*, pages 96–103, 2004.

[19] Naomi Nishimura, Prabhakar Ragde, and Stefan Szeider. Solving #SAT using vertex covers. *Acta Informatica*, **44**(7-8):509–523, 2007.

[20] Marko Samer and Stefan Szeider. Backdoor Trees. In *Proceedings of AAAI 08, Twenty-Third Conference on Artificial Intelligence, Chicago, Illinois, July 13–17, 2008*, pages 363–368, AAAI Press, 2008.

[21] Stefan Szeider. Minimal unsatisfiable formulas with bounded clause-variable difference are fixed-parameter tractable. *Journal of Computer and System Sciences*, **69**(4):656–674, 2004.

[22] Stefan Szeider. Generalizations of matched CNF formulas. *Annals of Mathematics and Artificial Intelligence*, **43**(1-4):223–238, 2005.

[23] Craig A. Tovey. A simplified NP-complete satisfiability problem. *Discrete Applied Mathematics*, **8**(1):85–89, 1984.

[24] Ryan Williams, Carla Gomes, and Bart Selman. On the connections between backdoors, restarts, and heavy-tailedness in combinatorial search. In *Informal Proc. of the Sixth International Conference on Theory and Applications of Satisfiability Testing, S. Margherita Ligure - Portofino, Italy, May 5-8, 2003 (SAT 2003)*, pages 222–230, 2003.