Algorithmica (2009) 54: 472–489
DOI 10.1007/s00453-008-9206-y

Improved Parameterized Set Splitting Algorithms: A Probabilistic Approach

Jianer Chen · Songjian Lu

Received: 8 September 2007 / Accepted: 9 June 2008 / Published online: 28 June 2008

Abstract In this paper, we study parameterized algorithms for the SET SPLITTING problem, for both weighted and unweighted versions. First, we develop a new and effective technique based on a probabilistic method that allows us to develop a simpler and more efficient deterministic kernelization algorithm for the unweighted SET SPLITTING problem. We then propose a randomized algorithm for the weighted SET SPLITTING problem that is based on a new subset partition technique and has its running time bounded by $O^*(2^k)$, which is significantly better than that of the previous best deterministic algorithm (which only works for the simpler unweighted SET SPLITTING problem) of running time $O^*(2.65^k)$. We also show that our algorithm can be de-randomized, which leads to a deterministic parameterized algorithm of running time $O^*(4^k)$ for the weighted SET SPLITTING problem and gives the first proof that the problem is fixed-parameter tractable.

Keywords Set splitting · Randomized algorithm · Derandomization · Parametrized algorithm

1 Introduction

Let X be a set. A *partition* of X is a pair of subsets (X_1, X_2) of X such that $X_1 \cup X_2 = X$ and $X_1 \cap X_2 = \emptyset$. We say that a subset S of X is *split* by the partition (X_1, X_2) of

A preliminary version of this paper was presented at The 13th Annual International Computing and Combinatorics Conference (COCOON 2007), Banff, Canada, July 2007, LNCS vol. 4598, pp. 537–547.

This work was supported in part by the National Science Foundation under the Grant CCF-0430683.

J. Chen · S. Lu (✉)
Department of Computer Science, Texas A&M University, College Station, TX 77843, USA
e-mail: sjlu@cs.tamu.edu

J. Chen
e-mail: chen@cs.tamu.edu

X if $S \cap X_1 \neq \emptyset$ and $S \cap X_2 \neq \emptyset$. The SET SPLITTING problem is defined as follows: given a collection $\mathcal{F}$ of subsets of a ground set X, construct a partition of X that maximizes the number of split subsets in $\mathcal{F}$.

A more generalized version of the SET SPLITTING problem is the *weighted* SET SPLITTING problem, in which each subset in the collection $\mathcal{F}$ is associated with a weight that is a real number, and the objective is to construct a partition of the ground set that maximizes the sum of the weights of the split subsets.

The SET SPLITTING problem is an important NP-hard problem [11]. A number of well-known NP-complete problems are related to the SET SPLITTING problem, including the HITTING SET problem that is to find a small subset of the ground X that intersects all subsets in a given collection $\mathcal{F}$, and the SET PACKING problem that is to find a large sub-collection $\mathcal{F}'$ of a given collection $\mathcal{F}$ of subsets such that the subsets in $\mathcal{F}'$ are all pairwise disjoint.

In terms of approximability, the SET SPLITTING problem is APX-complete [3]. Andersson and Engebretsen [2] gave a polynomial time approximation algorithm for the problem that has an approximation ratio bounded by 0.724. Zhang and Ling [17] presented an improved polynomial time approximation algorithm of approximation ratio 0.7499 for the problem. Better polynomial time approximation algorithms can be achieved if we further restrict the number of elements in each subset in the input [13, 17–19].

On certain applications, such as the analysis of micro-array data, people have studied the parameterized version of the SET SPLITTING problem, by associating each instance of the problem with a parameter k, which is in general a small positive integer [7]. The *parameterized unweighted* SET SPLITTING problem is defined as follows: given a triple $(X, \mathcal{F}, k)$, where X is a finite ground set, $\mathcal{F}$ is a collection of subsets of the ground set X, and k is the parameter that is a non-negative integer, decide if there is a partition of the ground set X that splits at lease k subsets in $\mathcal{F}$.

In this paper, we are mainly concerned with *parameterized algorithms* [9] for the parameterized SET SPLITTING problem, where the algorithms run in time $f(k)n^{O(1)}$, with $f(k)$ being a function that only depends on the parameter k. In particular, for small values of the parameter k, such an algorithm for the parameterized SET SPLITTING problem will become efficient.[1]

The parameterized unweighted SET SPLITTING problem has been studied in the literature. Dehne, Fellows, and Rosamond [7] were the first to study the problem and provided a parameterized algorithm of running time $O^*(72^k)$ for the problem.[2] In the same paper, the authors also proved that the parameterized unweighted SET SPLITTING problem has a kernel of fewer than $2k$ subsets: that is, there is a polynomial time algorithm that on a given instance $(X, \mathcal{F}, k)$ of parameterized unweighted SET SPLITTING, produces another instance $(X', \mathcal{F}', k')$ for the problem such that $|X'| \leq 4k^2$, $|\mathcal{F}'| < 2k$, $k' \leq k$, and that the set X has a partition that splits k subsets

[1] According to Downey and Fellows [9], a parameterized problem is *fixed parameter tractable* if it can be solved in time $f(k)n^{O(1)}$, where f is a function depending on the parameter k but independent of the input size n.

[2] Following the recent convention, by the notation $O^*(c^k)$, where $c > 1$ is a constant, we refer to a function of order $O(c^k n^{O(1)} g(k))$, where $g(k) = c^{o(k)}$.

in the collection $\mathcal{F}$ if and only if the set X' has a partition that splits k' subsets in the collection $\mathcal{F}'$. Later, Dehne, Fellows, Rosamond, and Shaw [8] developed an improved algorithm of running time $O^*(8^k)$ for the problem. The improved algorithm was obtained by combining the recently developed techniques of *greedy localization* and *modeled crown reduction* in the study of parameterized algorithms. The current best algorithm for the parameterized unweighted SET SPLITTING problem is developed by Lokshtanov and Sloper [15], where they used Chen and Kanj's result for the MAX-SAT problem [4] and reached a time complexity of $O^*(2.65^k)$.

A natural generalization of the parameterized unweighted SET SPLITTING problem is the *parameterized weighted* SET SPLITTING problem defined as follows: given a triple $(X, \mathcal{F}, k)$, where X is a finite ground set, $\mathcal{F}$ is a collection of subsets of the ground set X, in which each subset is assigned a weight (that is a real number), and k is the parameter that is a non-negative integer, either construct a partition of X that maximizes the weighted sum of k split subsets in $\mathcal{F}$, or report that no partition of X can split k subsets in $\mathcal{F}$. Note that there is an essential difference between parameterized unweighted SET SPLITTING and parameterized weighted SET SPLITTING. Parameterized unweighted SET SPLITTING is a decision problem that only requires a yes/no answer, while parameterized weighted SET SPLITTING is an optimization problem that, in case a partition of the ground set X splitting k subsets in $\mathcal{F}$ exists, requires to construct such a partition that maximizes the weighted sum of the split subsets.

No parameterized algorithms of running time of the form $f(k)n^{O(1)}$ have been known for the parameterized weighted SET SPLITTING problem. In fact, none of the techniques developed previously for the parameterized unweighted SET SPLITTING problem, such as those in [7, 8, 15], seems to be extendable to the weighted case.

In this paper, we develop new techniques in dealing with the SET SPLITTING problems for both weighted and unweighted cases. First, we develop a new and effective technique based on a probabilistic method that allows us to develop a *deterministic* kernelization algorithm for the parameterized unweighted SET SPLITTING problem. The new kernelization algorithm is simpler and more efficient compared with the previous kernelization algorithm given in [7]. We then propose a randomized algorithm for the parameterized weighted SET SPLITTING problem (thus, also for the parameterized unweighted SET SPLITTING problem) that is based on a new subset partition technique and has its running time bounded by $O^*(2^k)$. The running time of our randomized algorithm is significantly better than that of the previous best deterministic algorithm of running time $O^*(2.65^k)$ given in [15], which only works for the (simpler) parameterized unweighted SET SPLITTING problem. We also show that, using the structure of (n, k)-universal sets developed by Naor, Schulman, and Srinivasan [16], we can de-randomize our randomized algorithm, which leads to a parameterized algorithm of running time $O^*(4^k)$ for the weighted SET SPLITTING problem and gives the first proof that the problem is fixed parameter tractable.

A preliminary version of the current paper appeared in the proceedings of COCOON'07 [5], where, because of the page limit, some of the details were omitted. The current paper is a careful revision of [5] with all necessary details provided.

2 A New Kernelization Algorithm for SET SPLITTING

Since we will be only considering the parameterized versions of the SET SPLITTING problems, we will drop the word "parameterized unweighted" when we refer to the parameterized unweighted SET SPLITTING problem and drop the word "parameterized" when we refer to the parameterized weighted SET SPLITTING problem.

In this section, we focus on the SET SPLITTING problem. By a *kernelization algorithm* for SET SPLITTING, we mean a polynomial time algorithm that, on an instance $(X, \mathcal{F}, k)$ of SET SPLITTING, produces another instance $(X', \mathcal{F}', k')$ for the problem such that $k' \leq k$ and the size of the instance $(X', \mathcal{F}', k')$ only depends on the parameter k. The instance $(X', \mathcal{F}', k')$ will be called a *kernel* for the instance $(X, \mathcal{F}, k)$. Dehne, Fellows, and Rosamond [7] developed a kernelization algorithm by which the kernel $(X', \mathcal{F}', k')$ satisfies the conditions $|\mathcal{F}'| < 2k$ and that each subset in $\mathcal{F}'$ has at most $2k$ elements. Lokshtanov and Sloper [15] used the crown decomposition method to obtain a kernel such that both $|\mathcal{F}'|$ and $|X'|$ are less than $2k$. We introduce a new method to find the kernel for the SET SPLITTING problem. What is interesting in our method is that we use a probabilistic method to derive a deterministic kernelization algorithm. In particular, our method is simpler, has lower time complexity, and can also obtain a better kernel in term of the number of subsets in $\mathcal{F}'$ if there are subsets in $\mathcal{F}'$ whose size is larger than 2.

Lemma 2.1 *Given an instance $(X, \mathcal{F}, k)$ of the* SET SPLITTING *problem, let m_1 be the number of subsets in $\mathcal{F}$ that have only one element. If $|\mathcal{F}| - m_1 \geq 2k$, then a partition of X exists that splits at least k subsets in $\mathcal{F}$.*

Proof For each subset $S \in \mathcal{F}$, if S has at least two elements, we pick any two elements from S. Let V be the set of all these elements picked from the subsets in $\mathcal{F}$ that have more than one element. Note that for two subsets S_1 and S_2 in $\mathcal{F}$ that have more than one element, the two elements in S_1 and the two elements in S_2 may not be disjoint.

Suppose $|V| = t$. We randomly partition V into two subsets V_l and V_r, such that $|V_l| = \lfloor t/2 \rfloor$, $|V_r| = t - |V_l|$, i.e. we randomly pick $\lfloor t/2 \rfloor$ elements of V and put them in V_l and let the remaining $t - \lfloor t/2 \rfloor$ elements of V be in V_r. Thus, for any subset S in $\mathcal{F}$:

$$\Pr(S \text{ is split}) \begin{cases} \geq \frac{2\binom{t-2}{\lfloor t/2 \rfloor - 1}}{\binom{t}{\lfloor t/2 \rfloor}} = \frac{2\lfloor t/2 \rfloor (t - \lfloor t/2 \rfloor)}{t(t-1)} > \frac{1}{2}, \\ \quad \text{if } S \text{ has more than one element} \\ = 0, \quad \text{otherwise.} \end{cases}$$

If we let:

$$X_S = \begin{cases} 1, & \text{if } S \text{ is split,} \\ 0, & \text{otherwise,} \end{cases}$$

then the expectation of the number of split subsets in $\mathcal{F}$ satisfies

$$E\left(\sum_{S \in \mathcal{F}} X_S\right) \geq \frac{1}{2}(|\mathcal{F}| - m_1).$$

Therefore, if $|\mathcal{F}| - m_1 \geq 2k$, then the expectation of the number of split subsets in $\mathcal{F}$ is larger than or equal to k. That is, there must exist a partition of the ground set X such that the number of split subsets in $\mathcal{F}$ is at least k. This completes the proof of the lemma. □

The result of Lemma 2.1 was first observed by Lokshtanov and Sloper, who presented a proof in [15]. Our proof above is very different from that given in [15] and takes a probabilistic approach. Furthermore this new approach can lead to a better result for the kernelization when many subsets in $\mathcal{F}$ have more than 2 elements, as described in Lemma 2.4.

The following lemma shows that we can directly include subsets of at least k elements in our split subsets while we are solving the SET SPLITTING problem.

Lemma 2.2 *Let $(X, \mathcal{F}, k)$ be an instance of the* SET SPLITTING *problem, and let S be a subset in $\mathcal{F}$ that contains at least k elements. Then there is a partition of X that splits k subsets in $\mathcal{F}$ if and only if there is a partition of X that splits $k-1$ subsets in $\mathcal{F} - \{S\}$.*

Proof Suppose that there is a partition (X_l, X_r) of the ground set X that splits k subsets in $\mathcal{F}$. Then it is obvious that (X_l, X_r) splits (at least) $k-1$ subsets in $\mathcal{F} - \{S\}$.

On the other hand, suppose that there is a partition (X_l, X_r) of the ground set X that splits $k-1$ subsets $S_1, \ldots, S_{k-1}$ in $\mathcal{F} - \{S\}$. Let $l_i, r_i \in S_i$, $l_i \in X_l$, and $r_i \in X_r$, for all $1 \leq i \leq k-1$. Since S has at least k elements, there are at least two different elements l and r in S such that $l \notin \{r_1, \ldots, r_{k-1}\}$ and $r \notin \{l_1, \ldots, l_{k-1}\}$. Therefore, if we modify the partition (X_l, X_r) to enforce l in X_l and r in X_r (note that this modification still keeps l_i in X_l and r_i in X_r for all $1 \leq i \leq k-1$), then the new partition of X splits the subset S, as well as the $k-1$ subsets $S_1, \ldots, S_{k-1}$ in $\mathcal{F} - \{S\}$. In consequence, the new partition of the ground set X splits (at least) k subsets in the collection $\mathcal{F}$. □

Now we are ready to state our first kernelization result. For a given instance $(X, \mathcal{F}, k)$ of the SET SPLITTING problem, consider the following reduction rules.

Rule R1. If a subset S in $\mathcal{F}$ has only one element, remove S from $\mathcal{F}$.
Rule R2. If a subset S in $\mathcal{F}$ has at lease k elements, remove S from $\mathcal{F}$ and decrease k by 1.

The correctness of Rule **R1** is obvious: a subset of a single element can never be split by any partition of the ground set X. The correctness of Rule **R2** follows from Lemma 2.2.

Theorem 2.3 *Given an instance $(X, \mathcal{F}, k)$ of the* SET SPLITTING *problem, we can construct a kernel $(X_1, \mathcal{F}_1, k_1)$ such that $|\mathcal{F}_1| < 2k_1$, $k_1 \leq k$, $|X_1| < 2k_1^2$, and that each subset in $\mathcal{F}_1$ has at most $k_1 - 1$ elements. The running time of this process is bounded by $O(N)$, where N is the input size in terms of $(X, \mathcal{F}, k)$.*

Proof For the given instance $(X, \mathcal{F}, k)$, we first apply Rule **R1** to remove all subsets that contain a single element. Then, we apply bucket-sort to sort in linear time the

remaining subsets in $\mathcal{F}$ in non-increasing order in terms of their sizes. Finally, we apply Rule **R2** on the subsets in order in the sorted list, and stop at a subset on which Rule **R2** is not applicable or when the parameter value k reaches 0. This process obviously takes time $O(N)$.

Suppose that the instance produced by the above process is $(X_1, \mathcal{F}_1, k_1)$. If $k_1 = 0$ or $|\mathcal{F}_1| \geq 2k_1$, then by Lemma 2.1 (note that $\mathcal{F}_1$ contains no subsets of one element), $(X_1, \mathcal{F}_1, k_1)$ (as well as the original instance $(X, \mathcal{F}, k)$) is a "Yes" instance. In this case, our algorithm returns a trivial "Yes" instance $(\{a, b\}, \{\{a, b\}\}, 1)$. Otherwise, the correctness of the reduction rules **R1** and **R2** ensure that $(X_1, \mathcal{F}_1, k_1)$ is a "Yes" instance if and only if $(X, \mathcal{F}, k)$ is a "Yes" instance for the SET SPLITTING problem. So our algorithm simply returns $(X_1, \mathcal{F}_1, k_1)$.

To see that the instance $(X_1, \mathcal{F}_1, k_1)$ satisfies the conditions in the lemma, first note that if a non-trivial instance is returned by the process, then we must have $|\mathcal{F}_1| < 2k_1$. Moreover, since the subsets in $\mathcal{F}_1$ are sorted in non-increasing order in terms of their sizes, and Rule **R2** is not applicable to the first subset in the list, no subset in $\mathcal{F}_1$ contains more than $k_1 - 1$ elements. In consequence, we also have $|X_1| < 2k_1^2$. This completes the proof of the lemma. □

Theorem 2.3 improves the time complexity of the kernelization algorithm given in [7], which takes time $O(N + n^4)$, as well as that given in [15], which takes time $O(N + n^2)$, where n is the maximum of $|\mathcal{F}|$ and $|X|$.

Intuitively, when we randomly partition X into (X_l, X_r) such that each element in X has a probability of $1/2$ to be assigned to X_l and a probability of $1/2$ to be assigned to X_r, larger subsets (i.e., subsets with more elements) will have a better chance to be split. The following lemma confirms this intuition. Thus, if the collection $\mathcal{F}$ contains many large subsets, then we can obtain a better kernel, or a kernel with fewer subsets.

Lemma 2.4 *Let $(X, \mathcal{F}, k)$ be an instance of the* SET SPLITTING *problem. Suppose that the number of subsets of i elements in $\mathcal{F}$ is m_i for $1 \leq i \leq k-1$, and that the number of subsets that have at least k elements is m'_k. If $\sum_{i=2}^{k-1} \frac{2^i-2}{2^i} m_i + m'_k \geq k$, then a partition of X exists that splits at least k subsets in $\mathcal{F}$.*

Proof Let $S_1, \ldots, S_{m'_k}$ be the subsets in $\mathcal{F}$ that have at least k elements and let $\mathcal{F}_{<k} = \mathcal{F} - \{S_1, \ldots, S_{m'_k}\}$.

We use a randomized process to partition X into (X_l, X_r) and let each element in X go to X_l with a probability of $1/2$ and go to X_r with a probability of $1/2$, then for any subset $S \in \mathcal{F}_{<k}$ that has i elements:

$$\Pr(S \text{ is split}) = \frac{2^i - 2}{2^i}.$$

If we let:

$$X_S = \begin{cases} 1, & \text{if } S \text{ is split,} \\ 0, & \text{otherwise,} \end{cases}$$

then the expectation of the number of split subsets in $\mathcal{F}_{<k}$ satisfies

$$E\Big(\sum_{S\in\mathcal{F}_{<k}} X_S\Big) = \sum_{i=1}^{k-1}\sum_{|S|=i} E(S \text{ is split}) = \sum_{i=1}^{k-1}\frac{2^i-2}{2^i}m_i.$$

So there exists a partition of X such that the number of subsets in $\mathcal{F}_{<k}$ that are split is at least $\sum_{i=1}^{k-1}\frac{2^i-2}{2^i}m_i$. Hence if $\sum_{i=1}^{k-1}\frac{2^i-2}{2^i}m_i \geq k-m'_k$, there must exist a partition of X such that $k-m'_k$ subsets in $\mathcal{F}_{<k}$ are split. By repeatedly using Lemma 2.2, there is a partition of X that splits $k-m'_k+1$ subsets in $\mathcal{F}_{<k}\cup\{S_1\}$; there is a partition of X that splits $k-m'_k+2$ subsets in $\mathcal{F}_{<k}\cup\{S_1,S_2\}$; and so on. In conclusion, there is a partition of X that splits k subsets in $\mathcal{F}_{<k}\cup\{S_1,\ldots,S_{m'_k}\}=\mathcal{F}$. □

Using the procedure that is similar to Theorem 2.3, but counting the number of subsets in $\mathcal{F}$ of different size and using the result of Lemma 2.4, we have the following theorem that is stronger than Theorem 2.3.

Theorem 2.5 *Given an instance $(X,\mathcal{F},k)$ of the* SET SPLITTING *problem, we can find a kernel $(X_1,\mathcal{F}_1,k_1)$ in time $O(N)$ such that $|\mathcal{F}_1|<2k_1-\sum_{i=3}^{k_1-1}\frac{2^{i-1}-2}{2^{i-1}}m_i$, that $k_1\leq k$, that each subset in $\mathcal{F}_1$ has at most k_1-1 elements, and that $|X_1|<2k_1^2$, where N is the input size in terms of $(X,\mathcal{F},k)$, and m_i is the number of subsets of i elements in $\mathcal{F}_1$, $1\leq i\leq k_1-1$.*

Proof We use the same procedure as the one presented in Theorem 2.3 to find the kernel. Let the resulting instance be $(X_1,\mathcal{F}_1,k_1)$. We also calculate the values m_i from $\mathcal{F}_1$, for $1\leq i\leq k_1-1$. If $\sum_{i=2}^{k_1-1}\frac{2^i-2}{2^i}m_i\geq k_1$ (note that no subset in $\mathcal{F}_1$ contains more than k_1-1 elements), then by Lemma 2.4, the instance $(X_1,\mathcal{F}_1,k_1)$ is a "Yes" instance so we return a trivial "Yes" instance. Otherwise, we have $\sum_{i=2}^{k_1-1}\frac{2^i-2}{2^i}m_i<k_1$, which gives $|\mathcal{F}_1|=\sum_{i=2}^{k_1-1}m_i<2k_1-\sum_{i=3}^{k_1-1}\frac{2^{i-1}-2}{2^{i-1}}m_i$. In this case, we obtain a kernel $(X_1,\mathcal{F}_1,k_1)$ such that $|\mathcal{F}_1|<2k_1-\sum_{i=3}^{k_1-1}\frac{2^{i-1}-2}{2^{i-1}}m_i$, that $k_1\leq k$, that each subset in $\mathcal{F}_1$ has at most k_1-1 elements, and that $|X_1|<2k_1^2$. □

3 A Randomized Algorithm for Weighted SET SPLITTING

For the SET SPLITTING problem, Lokshtanov and Sloper [15] have currently the best parameterized algorithm, whose running time is bounded by $O^*(2.65^k)$. Unfortunately, their method does not seem to be extendable to the weighted case, neither do the methods presented in [7, 8] for the unweighted case. In fact, no previous work is known that gives a parameterized algorithm of running time of the form $f(k)n^{O(1)}$ for the weighted SET SPLITTING problem.

In this section, we present a randomized algorithm to solve the weighted SET SPLITTING problem. Our basic idea is that if a given instance $(X,\mathcal{F},k)$ of the weighted SET SPLITTING problem has a partition of the ground set X that splits k subsets in the collection $\mathcal{F}$, then there exists a subset X' of at most $2k$ elements in X

such that a proper partition of the elements in X' can split at least k subsets in $\mathcal{F}$. If we use a randomized process to partition X into (X_l, X_r) and let each element in X go to X_l with a probability of $1/2$ and go to X_r with a probability of $1/2$, then the probability that the elements in X' are partitioned properly is at least $2/2^{2k}$. Thus, if we try $O(4^k)$ times of the randomized partitioning of the ground set X, we have a good chance to find the proper partition of X' if it exists. In fact, a more thorough analysis reveals that only $O(2^k)$ trials are needed in this randomized algorithm.

Algorithm-1 SetSplitting$(X, \mathcal{F}, k)$
input: A ground set X, a collection $\mathcal{F}$ of subsets of X, and an integer k
output: A partition (X_l, X_r) of X and k subsets in $\mathcal{F}$ that are split by (X_l, X_r), or report "no partition of X splits k subsets in $\mathcal{F}$".

1. $\mathcal{Q}_0 = \emptyset$;
2. **for** $i = 1$ **to** $10 \cdot 2^k$ **do**
2.1. randomly partition X into X_l and X_r such that each element in X has a probability $1/2$ in X_l and a probability $1/2$ in X_r;
2.2. let $\mathcal{Q}$ be the collection of subsets in $\mathcal{F}$ that are split by (X_l, X_r);
2.3. **if** $\mathcal{Q}$ contains at least k subsets **then** delete all but the k subsets of maximum weight in $\mathcal{Q}$;
2.4. **if** the weighted sum of subsets in $\mathcal{Q}$ is larger than that in $\mathcal{Q}_0$ **then** $\mathcal{Q}_0 = \mathcal{Q}$;
3. return $\mathcal{Q}_0$.

Fig. 1 Randomized algorithm for WEIGHTED SET SPLITTING problem

Theorem 3.1 *The weighted* SET SPLITTING *problem can be solved by a randomized algorithm of running time* $O(2^k N)$, *where* N *is the input size in terms of* $(X, \mathcal{F}, k)$.

Proof Let $(X, \mathcal{F}, k)$ be an instance of the weighted SET SPLITTING problem. Suppose that there is a partition of the ground set X that splits at least k subsets in the collection $\mathcal{F}$. Let (X_l, X_r) be a partition of the ground set X and let $S_1, \ldots, S_k$ be k subsets in the collection $\mathcal{F}$ that are split by the partition (X_l, X_r), such that the weighted sum of $S_1, \ldots, S_k$ is the maximum over all collections of k subsets in $\mathcal{F}$ that can be split by a partition of X. More specifically, let $(l_1, r_1), \ldots, (l_k, r_k)$ be k pairs of elements in the ground set X such that $l_i, r_i \in S_i$, $l_i \in X_l$, and $r_i \in X_r$ for all $1 \le i \le k$. Note that it is possible that $l_i = l_j$ or $r_i = r_j$ for some $i \ne j$. In consequence, each of the sets $\{l_1, \ldots, l_k\}$ and $\{r_1, \ldots, r_k\}$ may contain fewer than k elements.

We construct a graph $G = (V, E)$, where $V = \{l_1, l_2, \ldots, l_k\} \cup \{r_1, r_2, \ldots, r_k\}$ and $E = \{(l_i, r_i) \mid 1 \le i \le k\}$. It is obvious that G is a bipartite graph with the left vertex set $L = \{l_1, l_2, \ldots, l_k\}$ and the right vertex set $R = \{r_1, r_2, \ldots, r_k\}$. Suppose that the graph G has t connected components $C_1, \cdots, C_t$, where $C_i = (V_i, E_i)$, with $n_i = |V_i|$ and $m_i = |E_i|$, for $1 \le i \le t$. Then $n_i \le m_i + 1$ for $1 \le i \le t$ and $\sum_{i=1}^{t} m_i = k$. If we use a randomized process to partition X into (X_l, X_r) and let each element in X go to X_l with a probability of $1/2$ and go to X_r with a probability of $1/2$, then for each connected component C_i of the graph G, the probability that the vertex set V_i of C_i is properly partitioned, i.e., either $L \cap V_i \subseteq X_l$ and $R \cap V_i \subseteq X_R$, or $R \cap V_i \subseteq X_l$ and

$L \cap V_i \subseteq X_R$, is $2/2^{n_i}$. Therefore, the total probability that the vertex set V_i for every connected component C_i is properly partitioned, i.e., that the pair (l_i, r_i) intersects with both X_l and X_r for all $1 \le i \le k$, is not less than

$$\frac{2}{2^{n_1}} \cdot \frac{2}{2^{n_2}} \cdots \frac{2}{2^{n_t}} \ge \frac{2}{2^{m_1+1}} \cdot \frac{2}{2^{m_2+1}} \cdots \frac{2}{2^{m_t+1}} = \frac{2^t}{2^{\sum_{i=1}^{t} m_i + t}} = \frac{1}{2^k}.$$

The algorithm in Fig. 1 implements the above idea. By the above discussion, each random partition (X_l, X_r) constructed in step 2.1 has a probability of at least $1/2^k$ to split the k subsets $S_1, \ldots, S_k$ (recall that $S_1, \ldots, S_k$ are the k subsets in $\mathcal{F}$ whose weighted sum is the maximum over all collections of k subsets in $\mathcal{F}$ that are split by a partition of X). Since step 2 loops $10 \cdot 2^k$ times, with a probability of at least

$$1 - \left(1 - \frac{1}{2^k}\right)^{10 \cdot 2^k} \ge 99.99\%,$$

one partition (X_l, X_r) constructed by step 2.1 splits the k subsets $S_1, \ldots, S_k$. For this partition (X_l, X_r), steps 2.2–2.4 produces a collection $\mathcal{Q}$ of k subsets in $\mathcal{F}$ whose weighted sum is the maximum over all collections of k subsets in $\mathcal{F}$ that can be split by a partition of the ground set X.

Since each execution of steps 2.1–2.4 obviously takes time $O(N)$, we conclude that the running time of the algorithm **SetSplitting** is bounded by $O(2^k N)$.

For a general error bound $\epsilon > 0$, we can simply run the algorithm **SetSplitting** c times, where the constant c satisfies the condition $(1 - 0.9999)^c \le \epsilon$, which will produce, in time $O(2^k N)$ and with a probability at least $1 - \epsilon$, a collection $\mathcal{Q}$ of k subsets in $\mathcal{F}$ whose weighted sum is the maximum over all collections of k subsets in $\mathcal{F}$ that can be split by a partition of the ground set X. □

Obviously, the randomized algorithm **SetSplitting** of running time $O(2^k N)$ can be directly used to solve the simpler (unweighted) SET SPLITTING problem, and its running time is significantly better than that of the previous best deterministic algorithm [15] for the problem. Moreover, the algorithm **SetSplitting** is much simpler than the one presented in [15]. The algorithm in [15] needs to call the algorithm for the parameterized MAX-SAT problem developed in [4], which is quite involved.

By combining the kernelization algorithm, the time complexity for the SET SPLITTING problem can be further improved.

Theorem 3.2 *The* (*unweighted*) SET SPLITTING *problem can be solved by a randomized algorithm of running time* $O(2^k k^2 + N)$, *where* N *is the input size in terms of* $(X, \mathcal{F}, k)$.

4 Derandomization

The randomized algorithm in the previous section can be de-randomized via a deterministic construction of an (n, k)-*universal set* [16], which is described in this section. We will always assume that n and k are two integers such that $n \ge k$.

Denote by Z_n the set $\{0, 1, \ldots, n-1\}$. A *splitting function* over Z_n is a $\{0,1\}$ (i.e., Boolean) function over Z_n. A splitting function f over Z_n can be naturally represented as a binary string s of length n such that the i-th bit of s is 0 if and only if $f(i) = 0$. Moreover, a splitting function can be interpreted as a partition (X_1, X_2) of the set Z_n (i.e., putting all x in Z_n such that $f(x) = 0$ in X_1 and putting all y in Z_n such that $f(y) = 1$ in X_2).

A subset S of Z_n is a *k-subset* if S consists of exactly k elements. Let (S_1, S_2) be a partition of the k-subset S. We say that a splitting function f over Z_n *implements* the partition (S_1, S_2) of S if $f(x) = 0$ for all $x \in S_1$ and $f(y) = 1$ for all $y \in S_2$.

Definition 1 [16] A set $\mathcal{P}$ of splitting functions over Z_n is an (n,k)-*universal set* if for every k-subset S of Z_n and any partition (S_1, S_2) of S, there is a splitting function f in $\mathcal{P}$ that implements (S_1, S_2). The *size* of an (n,k)-universal set $\mathcal{P}$ is the number of splitting functions in $\mathcal{P}$.

The best known deterministic construction of (n,k)-universal sets was developed by Noar, Schulman, and Srinivasan, and was described via the construction of a more general structure, i.e., (n,k,l)-splitters. Moreover, the construction was presented in an extended abstract [16] in which many details were omitted. For the completeness of our discussion, we will re-produce in this section the constructions and analysis related to (n,k)-universal sets, and provide all needed details. We will also show in detail how these techniques are used to derive an efficient deterministic parameterized algorithm for the weighted SET SPLITTING problem.

We start with some terminologies and definitions in probability theory.

Let $(\Omega, \Pr)$ be a probability space, where Ω is a finite set and $\Pr$ is the probability measure. The *size* of $(\Omega, \Pr)$ is the number of elements in Ω. The probability space $(\Omega, \Pr)$ is *uniform* if $\Pr(a) = 1/|\Omega|$ for all $a \in \Omega$ (in this case, we will simply write the probability space as Ω).

A $\{0,1\}$-*random variable* ξ over the probability space $(\Omega, \Pr)$ is a function from Ω to $\{0,1\}$. A group of h $\{0,1\}$-random variables $\xi_1, \xi_2, \ldots, \xi_h$ are *mutually independent* if for any combination of h binary bits $b_1, \ldots, b_h$ in $\{0,1\}$, the following holds:

$$\Pr(\xi_1 = b_1, \xi_2 = b_2, \ldots, \xi_h = b_h) = \Pr(\xi_1 = b_1) \cdot \Pr(\xi_2 = b_2) \cdot \cdots \cdot \Pr(\xi_h = b_h).$$

A group of n $\{0,1\}$-random variables $\xi_1, \xi_2, \ldots, \xi_n$ are *k-wise independent* if every group of k different $\{0,1\}$-random variables among $\xi_1, \xi_2, \ldots, \xi_n$ are mutually independent.

The following lemma is crucial for our construction, and was first proved in [1].

Lemma 4.1 [1] *Let $n = 2^d - 1$ for an integer d and $k \le n$ be an odd number. There is an algorithm of running time $O(n(n+1)^{(k-1)/2})$ that constructs a uniform probability space Ω of size $2(n+1)^{(k-1)/2}$ and a group of n k-wise independent $\{0,1\}$-random variables $\xi_1, \ldots, \xi_n$ over Ω such that $\Pr(\xi_i = 0) = \Pr(\xi_i = 1) = 1/2$ for all $1 \le i \le n$.*

We need the following observation for our analysis.

Lemma 4.2 *Let $G = (V_1 \cup V_2, E)$ be a bipartite graph with the vertex bipartition (V_1, V_2), where $|V_1| = n$ and $|V_2| = m$. Let ρ be a real number, $0 \le \rho \le 1$. If every vertex in V_1 has degree at least ρm, then there is at least one vertex in V_2 whose degree is at least ρn.*

Proof Since each vertex in V_1 has degree at least ρm in the bipartite graph G, the total number of edges in G is at least ρmn. Therefore, among the m vertices in V_2, at least one of them has degree at least ρn. □

Now we present the construction of an (n, k)-universal set of small size that, however, is not sufficiently efficient. Compared with the work presented in [16], the size of our structure is more precise (and slightly improved), which will be important for the later construction. Moreover, the time complexity of our construction is more efficient than that described in [16].

Lemma 4.3 *Let k be an odd number, $k \le n$. There is an (n, k)-universal set of size bounded by $2ek2^k \log n$, which can be constructed in time $O(\binom{n}{k} k2^k (2n)^{(k-1)/2})$, where e is the base of the natural logarithm.*

Proof Let $n_1 = 2^d - 1$, where d is the smallest integer such that $n \le n_1$ (note that $n \le n_1 \le 2n - 1$). By Lemma 4.1, we can construct, in time $O(n_1(n_1 + 1)^{(k-1)/2})$, a uniform probability space Ω of size $2(n_1 + 1)^{(k-1)/2}$ and a group of n_1 k-wise independent $\{0, 1\}$-random variables $\xi_1, \ldots, \xi_{n_1}$ over Ω such that $\Pr(\xi_i = 0) = \Pr(\xi_i = 1) = 1/2$ for all $1 \le i \le n_1$. Pick the first n of these n_1 random variables, we get a group of n k-wise independent $\{0, 1\}$-random variables $\xi_1, \ldots, \xi_n$ over the uniform probability space Ω such that $\Pr(\xi_i = 0) = \Pr(\xi_i = 1) = 1/2$ for all $1 \le i \le n$. All these can be constructed in time $O(n(2n)^{(k-1)/2})$.

Note that the uniform probability space Ω and the random variables $\xi_1, \ldots, \xi_n$ constructed above actually make a collection $\mathcal{P}$ of $D = 2(n_1 + 1)^{(k-1)/2}$ splitting functions over Z_n. In fact, for each element a in Ω, the values of the random variables $\xi_1, \ldots, \xi_n$ make a binary string $\xi_1(a) \cdots \xi_n(a)$ of length n, which, as we explained above, can be interpreted as a splitting function over Z_n.

Construct a bipartite graph $G = (V_1 \cup V_2, E)$ with the vertex bipartition (V_1, V_2), where V_1 consists of D vertices, corresponding to the D splitting functions in the collection $\mathcal{P}$, and the vertex set V_2 consists of $D' = \binom{n}{k} 2^k$ vertices such that for each k-subset S of Z_n and each partition (S_1, S_2) of S, there is a corresponding vertex in V_2. An edge $[v, w]$ is created in G if the splitting function corresponding to the vertex $v \in V_1$ implements the partition (S_1, S_2) of a k-subset S of Z_n that correspond to the vertex $w \in V_2$.

Claim Each vertex in V_2 has a degree $D/2^k$.

To see why the claim holds, let $S = \{h_1, \ldots, h_k\}$ be any k-subset of Z_n and let (S_1, S_2) be a partition of S. Define k binary bits b_{h_i}, $1 \le i \le k$ such that $b_{h_i} = 0$ if $h_i \in S_1$ and $b_{h_i} = 1$ if $h_i \in S_2$. Consider the k mutually independent random variables

$\xi_{h_1}, \ldots, \xi_{h_k}$ (they are mutually independent because the random variables $\xi_1, \ldots, \xi_n$ are k-wise independent), we have

$$\Pr(\xi_{h_1} = b_{h_1}, \ldots, \xi_{h_k} = b_{h_k}) = \Pr(\xi_{h_1} = b_{h_1}) \cdots \Pr(\xi_{h_k} = b_{h_k}) = 1/2^k$$

That is, there are $D/2^k$ elements a in Ω such that $\xi_{h_i}(a) = b_{h_i}$ for all $1 \le i \le k$. Using the interpretation above, there are $D/2^k$ splitting functions in the collection $\mathcal{P}$ that implement the partition (S_1, S_2) of the k-subset S. By our construction of the graph G, the vertex w in V_2 corresponding to the partition (S_1, S_2) of the k-subset S has degree $D/2^k$. The claim now is proved because S is an arbitrary k-subset in Z_n and (S_1, S_2) is an arbitrary partition of S.

By Lemma 4.2, there is a vertex v_1 in V_1 in the graph G whose degree is at least $D'/2^k$. In other words, there is a splitting function in the collection $\mathcal{P}$ that implements at least $D'/2^k$ partitions of k-subsets of Z_n (these partitions can be partitions for different k-subsets of Z_n).

We perform the following operations on the graph G: mark the vertex v_1 in V_1 and remove all vertices in V_2 that are adjacent to v_1 (or, equivalently, we mark a splitting function f in $\mathcal{P}$ and remove all partitions of k-subsets of Z_n that are implemented by f). Let D'_1 be the number of vertices in V_2 in the remaining bipartite graph G'. $D'_1 \le (1 - 1/2^k)D'$.

Note that each vertex in V_2 in the remaining graph G' still has degree $D/2^k$. Therefore, by repeating the above process, in the remaining graph G', we can find a vertex v_2 in V_1 that is adjacent to at least $D'_1/2^k$ vertices in V_2. Now we mark v_2, and remove the vertices in V_2 that are adjacent to v_2. Now there are at most $D'_2 \le (1 - 1/2^k)D'_1 \le (1 - 1/2^k)^2 D'$ vertices in V_2 in the remaining graph.

Repeat the above process until all vertices in V_2 are removed. The number t' of times the above process is repeated is not larger than the smallest integer t such that $(1 - 1/2^k)^t D' < 1$. Recall that $D' = \binom{n}{k}2^k$, we get $t' \le ek2^k(\log n + 1)$.

Each execution of the above process marks a vertex in V_1, therefore, there are at most $ek2^k(\log n + 1)$ vertices in V_1 that are marked in the above process. By our construction, all vertices in the set V_2 are adjacent to at least one marked vertex in V_1. Accordingly, there are $D'' \le ek2^k(\log n + 1)$ splitting functions in the collection $\mathcal{P}$ such that every partition of any k-subset in Z_n is implemented by at least one of these D'' splitting functions. That is, these D'' splitting functions make an (n, k)-universal set $\mathcal{P}'$ of size bounded by $ek2^k(\log n + 1) \le 2ek2^k \log n$.

Now we analyze the time complexity for the entire construction of the (n, k)-universal set $\mathcal{P}'$. As given earlier, the construction of the uniform probability space Ω and the $\{0, 1\}$ random variables $\xi_1, \ldots, \xi_n$, takes time $O(n(2n)^{(k-1)/2})$. The construction of the bipartite graph G takes time $O(k|V_1||V_2|) = O(kDD') = O(\binom{n}{k}k2^k(2n)^{(k-1)/2})$. To perform the above iteration process of marking vertices in V_1, we can represent the bipartite graph G has a $D \times D'$ matrix, and keep an array for the degrees of the vertices in V_1. It is not difficult to verify that on such data structures, the entire vertex marking process takes time $O(t'D + DD') = O(\binom{n}{k}2^k(2n)^{(k-1)/2})$. Thus, the total construction of the (n, k)-universal set $\mathcal{P}'$ takes time $O(\binom{n}{k}2^k k(2n)^{(k-1)/2})$. □

The size of the (n,k)-universal set constructed in Lemma 4.3 is quite small. Unfortunately, the time $O(\binom{n}{k}2^k k(2n)^{(k-1)/2})$ for constructing such an (n,k)-universal set given in the lemma is unacceptably high. Therefore, we need to play further tricks to reduce the construction time.

Before we present our further construction, we give an intuitive explanation for the basic idea. The construction time $O(\binom{n}{k}2^k k(2n)^{(k-1)/2})$ in Lemma 4.3 is of order $n^{O(k)}$. Observing that $k^{k/\log k} = 2^k$, we try to (1) reduce the size of the ground set from n to $k^{O(1)}$, and (2) reduce the parameter value k to a value k_1 of order $k/\log k$. In order to do (1), we apply a construction described in [10], which produces a family $\mathcal{H}$ of $O(n)$ mappings from Z_n to Z_{k^2} such that for any k-subset S in Z_n there is a mapping in $\mathcal{H}$ that is injective from S to Z_{k^2}. This enables us to concentrate on (k^2,k)-universal sets because the composition of any (k^2,k)-universal set and the family $\mathcal{H}$ gives an (n,k)-universal set. To accomplish (2), we construct a family $\mathcal{B}$ of partitions of the ground set Z_{k^2} into $t = O(\log k)$ parts such that every k-subset S in Z_{k^2} is evenly distributed in the t parts for at least one partition in the family $\mathcal{B}$. Now for each partition $(X_1,\ldots,X_t)$ in the family $\mathcal{B}$, we construct the $(|X_i|,k/t)$-universal sets, for $i = 1,\ldots,t$. Since $|X_i| \le k^2$ and $k/t = O(k/\log k)$, these universal sets with the reduced ground set size and reduced parameter value can be constructed efficiently. Note that every partition of any k-subset in Z_{k^2} can be implemented by a combination of t splitting functions in the t universal sets for some partitions $(X_1,\ldots,X_t)$ of the ground set Z_{k^2}. We also carefully construct the family $\mathcal{B}$ so that the total number of partitions in $\mathcal{B}$ is bounded by $2^{o(k)}$. Combining all these constructions will give an (n,k)-universal set of size $O^*(2^k)$.

Formally, fix n and k, where $k \le n$. Define

$$\begin{aligned} k_1 &= \text{ the largest odd number bounded by } k/(4\log k),\\ t &= \lceil k/k_1 \rceil, \quad \text{(it is not hard to verify that } t \le 4\log k + 2),\\ k_2 &= k - k_1(t-1), \quad \text{(note that } k_2 \le k_1),\\ n_1 &= k^2,\\ p &= \text{ a prime number such that } n \le p < 2n, \end{aligned} \tag{1}$$

where the existence of the prime number p above is guaranteed by Bertrand's Conjecture [12].

Consider the set $Z_{k^2} = \{0,1,\ldots,k^2-1\}$. Pick any $t-1$ elements $i_2, i_3, \ldots, i_t$ in Z_{k^2}, such that $i_2 < i_3 < \cdots < i_t$. These $t-1$ elements naturally divide the set Z_{k^2} into t sets consisting of consecutive elements (where X_1 may be an empty set):

$$X_1 = \{0,\ldots,i_2-1\}, \qquad X_2 = \{i_2,\ldots,i_3-1\}, \ldots, X_t = \{i_t,\ldots,k^2-1\}.$$

Such a division $(X_1, X_2, \ldots, X_t)$ of the set Z_{k^2} based on $t-1$ selected elements in Z_{k^2} will be called a *t-grouping* of the set Z_{k^2}.

According to Lemma 4.3, we can construct, noting that k_1 is an odd number, an (n_1,k_1)-universal set $\mathcal{P}_1$ of size $D_1 \le 2ek_1 2^{k_1}\log n_1$ in time $O(\binom{n_1}{k_1}2^{k_1}k_1(2n_1)^{(k_1-1)/2})$. Moreover, we define $k_2' = k_2$ if k_2 is odd, and $k_2' = k_2+1$ if k_2 is even, and construct an (n_1,k_2')-universal set $\mathcal{P}_2$ of size $D_2 \le 2ek_2' 2^{k_2'}\log n_1 \le$

$2e(k_2+1)2^{k_2+1}\log n_1$ in time $O(\binom{n_1}{k_1}2^{k_1}k_1(2n_1)^{(k_1-1)/2})$ (we replaced k_2' by k_1 in the time complexity simply because $k_2' \le k_1$). Using the definitions of k_1, k_2, and n_1, it is not hard to verify that

$$D_1 \le 4k2^{k_1}, \quad \text{and} \quad D_2 \le 4k2^{k_2+1}. \tag{2}$$

Lemma 4.4 *Let $\mathcal{P}$ be an (n,k)-universal set. Then for any n', $k \le n' \le n$, $\mathcal{P}$ is also an (n',k)-universal set, and for any $k' \le k$, $\mathcal{P}$ is also an (n,k')-universal set.*

Proof Each splitting function in $\mathcal{P}$ can be regarded as a splitting function over $Z_{n'}$. Since every k-subset of $Z_{n'}$ is also a k-subset of Z_n, we conclude that any partition of any k-subset in $Z_{n'}$ is implemented by a splitting function in $\mathcal{P}$, i.e., $\mathcal{P}$ is also an (n',k)-universal set.

Every partition (S_1', S_2') of any k'-subset S' of Z_n can be extended to a partition (S_1, S_2) of a k-subset of Z_n by adding $k-k'$ elements in $Z_n - S$ to S_1'. Now the splitting function in $\mathcal{P}$ that implements (S_1, S_2) also implements the partition (S_1', S_2') of S'. Thus, $\mathcal{P}$ is also an (n,k')-universal set. □

Now we are ready to construct our (n,k)-universal set $\mathcal{P}$. Each splitting function over Z_n in $\mathcal{P}$ is defined based on an integer z between 0 and $p-1$, a t-grouping $(X_1,\ldots,X_t)$ of the set Z_{k^2}, $t-1$ splitting functions $f_1,\ldots,f_{t-1}$ in the (n_1,k_1)-universal set $\mathcal{P}_1$, and a splitting function f_t in the (n_1,k_2')-universal set $\mathcal{P}_2$. The splitting function over Z_n is defined in Fig. 2.

First note that the function $f_{z,(X_1,\ldots,X_t),(f_1,\ldots,f_{t-1},f_t)}(a)$ is a well-defined splitting function. In fact, by step 1, x is an element in Z_{k^2}. Since $(X_1,\ldots,X_t)$ is a t-grouping of Z_{k^2}, x must belong to a unique X_i and have a unique rank j in X_i. Thus, step 3 will return a Boolean value $f_i(j)$.

Definition Let $\mathcal{P}$ be the collection of all possible splitting functions over Z_n defined in Fig. 2, over all integers z, $0 \le z < p$, all t-groupings $(X_1,\ldots,X_t)$ of Z_{k^2}, all possible lists $(f_1,\ldots,f_{t-1})$ of splitting functions in the (n_1,k_1)-universal set $\mathcal{P}_1$ (where the same function may appear more than once in the list), and all splitting functions f_t in the (n_1,k_2')-universal set $\mathcal{P}_2$.

We will show that the collection $\mathcal{P}$ is the desired (n,k)-universal set. For this, we still need one more lemma. We say that a function f on Z_n is *injective from a subset S of Z_n* if for any two different elements x and y in S, $f(x) \ne f(y)$.

Splitting $f_{z,(X_1,\ldots,X_t),(f_1,\ldots,f_{t-1},f_t)}(a)$
where $a \in Z_n$ is the input of the function, $0 \le z < p$, $(X_1,\ldots,X_t)$ is a t-grouping of Z_{k^2}, $f_1,\ldots,f_{t-1}$ are splitting functions in $\mathcal{P}_1$, and f_t is a splitting function in $\mathcal{P}_2$.
1. $x = (az \bmod p) \bmod k^2$;
2. suppose that x is the j-th smallest element in X_i;
3. return $f_i(j)$.

Fig. 2 A splitting function over Z_n

Lemma 4.5 [10] *Let p be a prime such that $n \le p < 2n$, and let S be a k-subset in Z_n. Then there is an integer z, $0 \le z < p$, such that the function g_z over Z_n, defined as $g_z(a) = (az \bmod p) \bmod k^2$, is injective from S.*

Now we are ready for our main result in this section.

Theorem 4.6 [16] *The collection $\mathcal{P}$ defined above is an (n,k)-universal set of size bounded by $n2^{k+12\log^2 k+12\log k+6} = n2^{k+o(k)}$, and can be constructed in time $O(n2^{k+12\log^2 k+12\log k}) = O(n2^{k+o(k)})$.*

Proof We first consider the size of $\mathcal{P}$. There are $p < 2n$ possible integers z. As described earlier, each t-grouping of Z_{k^2} can be given by $t-1$ different elements in Z_{k^2}. Therefore, the total number of different t-groupings of Z_{k^2} is bounded by $\binom{k^2}{t-1} \le k^{2(t-1)}$. The number of possible lists of $(f_1, \ldots, f_{t-1})$ of splitting functions in $\mathcal{P}_1$ is $|\mathcal{P}_1|^{t-1} = D_1^{t-1}$, and finally, the number of splitting functions in $\mathcal{P}_2$ is $|\mathcal{P}_2| = D_2$. Putting all these together, and recall the definitions and inequalities in (1) and (2), we conclude that the size of $\mathcal{P}$ is bounded by $2nk^{2(t-1)}D_1^{t-1}D_2 \le n2^{k+12\log^2 k+12\log k+6}$.

To construct the collection $\mathcal{P}$, we first construct the collections $\mathcal{P}_1$ and $\mathcal{P}_2$. As discussed earlier, these two collections can be constructed in time $O(\binom{n_1}{k_1}2^{k_1}k_1(2n_1)^{(k_1-1)/2}) = O(2^k)$. Once the collections $\mathcal{P}_1$ and $\mathcal{P}_2$ are available, the integers z, the t-groupings $(X_1, \ldots, X_t)$ of Z_{k^2}, and the lists $(f_1, \ldots, f_{t-1})$ of splitting functions in $\mathcal{P}_1$ and the functions f_t in $\mathcal{P}_2$ can be systematically enumerated, in constant time per combination, which gives a representation of the corresponding splitting function in $\mathcal{P}$. In conclusion, the collection $\mathcal{P}$ can be constructed in time $O(|\mathcal{P}|) = O(n2^{k+12\log^2 k+12\log k})$.

What remains is to show that $\mathcal{P}$ is an (n,k)-universal set. For this, let S be a given k-subset of Z_n and let (S_1, S_2) be a partition of S. By Lemma 4.5, there is an integer z_0, $0 \le z_0 < p$, such that the function g_{z_0} over Z_n is injective from S. Let S', S'_1, and S'_2 be the subsets of Z_{k^2} that are the images of S, S_1, and S_2 under g_{z_0}, respectively. By the definitions, we have $|S'| = |S|$, $|S'_1| = |S_1|$, $|S'_2| = |S_2|$, and (S'_1, S'_2) is a partition of the k-subset S' in Z_{k^2}.

It is easy to see that there is a t-grouping $(X_1^0, \ldots, X_t^0)$ of the set Z_{k^2} such that each of the first $t-1$ subsets $X_1^0, \ldots, X_{t-1}^0$ contains exactly k_1 elements in S', and the last subset X_t^0 contains k_2 elements in S'. Let $T_i = X_i^0 \cap S'$ for $1 \le i \le t$. Then T_i is a k_1-subset of X_i^0 for $1 \le i \le t-1$, and T_t is a k_2-subset of X_t^0. Moreover, the partition (S'_1, S'_2) of S' induces a partition $(T_{i,1}, T_{i,2})$ for each T_i, $1 \le i \le t$, where $T_{i,1} = T_i \cap S'_1$ and $T_{i,2} = T_i \cap S'_2$.

Since $\mathcal{P}_1$ is an (n_1, k_1)-universal set, which by Lemma 4.4 is also a $(|X_i^0|, k_1)$-universal set, for each i, $1 \le i \le t-1$, there is a splitting function f_i^0 in $\mathcal{P}_1$ that implements the partition $(T_{i,1}, T_{i,2})$ of the k_1-subset T_i of X_i^0 (note that the subset X_i^0 can be regarded as the set $Z_{|X_i^0|}$), for $1 \le i \le t-1$. That is, $f_i^0(x) = 0$ if $x \in T_{i,1}$ and $f_i^0(y) = 1$ if $y \in T_{i,2}$. Similarly, there is a splitting function f_t^0 in $\mathcal{P}_2$ that implements the partition $(T_{t,1}, T_{t,2})$ of the k_2-subset T_t.

Now consider the splitting function $f_{z_0,(X_1^0,\ldots,X_t^0),(f_1^0,\ldots,f_t^0)}$. On an element a in the subset S_1, step 1 of the algorithm **Splitting** produces an element $x = g_{z_0}(a)$ in the set S_1'. Suppose that x is in the set X_i^0, then x is in the set $T_{i,1}$. By the way we selected the splitting function f_i^0, we have $f_i^0(x) = 0$. In summary, on an element a in the subset S_1, we have $f_{z_0,(X_1^0,\ldots,X_t^0),(f_1^0,\ldots,f_t^0)}(a) = 0$. Using exactly the same reasoning, we can show $f_{z_0,(X_1^0,\ldots,X_t^0),(f_1^0,\ldots,f_t^0)}(a) = 1$ for every element a in S_2. Therefore, the function $f_{z_0,(X_1^0,\ldots,X_t^0),(f_1^0,\ldots,f_t^0)}$ in the collection $\mathcal{P}$ implements the partition (S_1, S_2) of the k-subset S of Z_n.

Since S is an arbitrary k-subset of Z_n and (S_1, S_2) is an arbitrary partition of S, we conclude that the collection $\mathcal{P}$ is an (n, k)-universal set. □

Using Theorem 4.6, we can derandomize our algorithm in the last section, and obtain a deterministic parameterized algorithm of running time $O^*(4^k)$ for the weighted SET SPLITTING problem. This also provides the first proof for the fixed parameter tractability of the problem.

Theorem 4.7 *The weighted* SET SPLITTING *problem can be solved by a deterministic algorithm of running time* $O(N^2 4^{k+6\log^2 k+6\log k}) = O(N^2 4^{k+o(k)})$, *where* N *is the instance size of the problem.*

Proof Let $(X, \mathcal{F}, k)$ be an instance of the weighted SET SPLITTING problem, where without loss of generality, let the ground set X be Z_n. We construct an $(n, 2k)$-universal set $\mathcal{P}$ based on Theorem 4.6, in time $O(n2^{2k+12\log^2(2k)+12\log(2k)}) = O(n4^{k+o(k)})$.

We use each splitting function in $\mathcal{P}$ to partition the ground set X, and see if the corresponding partition of X splits at least k subsets in $\mathcal{F}$. If so, we record the collection of the k subsets of the largest weight that are split by this partition. We repeat this process for all splitting functions in $\mathcal{P}$. The output of our algorithm is either "No" if no partition of X is constructed in this process that splits k subsets in $\mathcal{F}$, or the collection of the k subsets of the largest weight over all collections recorded in this process.

If the answer to the instance $(X, \mathcal{F}, k)$ is "No", then the above algorithm obviously returns "No" because the algorithm does not return "No" only if it actually constructs a partition of X that splits k subsets in $\mathcal{F}$. On the other hand, if the answer to the instance is not "No", then there is a partition of X that splits k subsets $S_1, \ldots, S_k$ in $\mathcal{F}$ whose total weight is the largest over all k split subsets caused by partitions of X. As we explained in the previous section, there is a set W of at most $2k$ elements in X and a partition (W_1, W_2) of W such that $W_1 \cap S_i \neq \emptyset$ and $W_2 \cap S_i \neq \emptyset$, for all $1 \leq i \leq k$. Therefore, when we perform the above process using a splitting function f in $\mathcal{P}$ that implements (W_1, W_2) (note by Lemma 4.4, f is an $(n, |W|)$-universal set even if $|W| < 2k$), the corresponding partition of X will split all these k subsets $S_1, \ldots, S_k$, and record the collection of k subsets of largest weight in $\mathcal{F}$ that can be split by a partition of the ground set X. □

5 Conclusion

In this paper, we studied parameterized algorithms for the SET SPLITTING problem. We developed a new and effective technique based on a probabilistic method that allows us to develop a simpler and more efficient (deterministic) kernelization algorithm for the unweighted SET SPLITTING problem. This new technique has also led to a better kernel if many subsets in the input have more than two elements.

We proposed a randomized algorithm for the weighted SET SPLITTING problem that is based on a new subset partition technique and has its running time bounded by $O^*(2^k)$. This is significantly better than the previous best known upper bound (which is a deterministic algorithm that only works for the simpler unweigthed SET SPLITTING problem). We also showed that our algorithm can be de-randomized, thus providing the first proof for the fixed parameter tractability of the weighted SET SPLITTING problem.

The de-randomization process based on the construction of (n, k)-universal sets is of general interest and has been used recently in the development of efficient parameterized algorithms for other problems, such as k-PATH, MATCHING and PACKING problems that can be solved by randomized divide-and-conquer methods [6, 14].

We note that in the discussion of Theorem 2.5 on the kernelization algorithm for the unweighted SET SPLITTING problem, if the instance $(X_1, \mathcal{F}_1, k_1)$ satisfies the condition $|\mathcal{F}_1| \geq 2k_1 - \sum_{i=3}^{k_1-1} \frac{2^{i-1}-2}{2^{i-1}} m_i$, where m_i is the number of subsets of i elements in $\mathcal{F}_1$, $1 \leq i \leq k_1 - 1$, we can directly conclude that $(X_1, \mathcal{F}_1, k_1)$ is a "Yes" instance without providing an actual partition of the ground set X_1 that splits k_1 subsets in $\mathcal{F}_1$. Our probabilistic analysis given in Lemma 2.4 also does not seem to hint an easy construction of such a partition. It will be interesting to see whether such a partition can be constructed efficiently in this case.

References

1. Alon, N., Babai, L., Itai, A.: A fast and simple randomized parallel algorithm for the maximal independent set problem. J. Algorithms **7**, 567–683 (1986)
2. Andersson, G., Engebretsen, L.: Better approximation algorithms and tighter analysis for set splitting and not-all-equal Sat. In: ECCCTR: Electronic Colloquium on Computational Complexity (1997)
3. Ausiello, G., Crescenzi, P., Gambosi, G., Kann, V., Marchetti-Spaccamela, A., Protasi, M.: Complexity and Approximation: Combinatorial Optimization Problems and Their Approximability Properties. Springer, Berlin (1999)
4. Chen, J., Kanj, I.: Improved exact algorithms for Max-Sat. Discrete Appl. Math. **142**, 17–27 (2004)
5. Chen, J., Lu, S.: Improved algorithm for weighted and unweighted set splitting problems. In: COCOON 2007. Lecture Notes in Computer Science, vol. 4598, pp. 573–547 (2007)
6. Chen, J., Lu, S., Sze, S., Zhang, F.: Improved algorithms for path, matching, and packing problems. In: Proc. of the Eighteen Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2007), pp. 298–307 (2007)
7. Dehne, F., Fellows, M., Rosamond, F.: An FPT algorithm for set splitting. In: WG 2003. Lecture Notes in Computer Science, vol. 2880, pp. 180–191 (2003)
8. Dehne, F., Fellows, M., Rosamond, F., Shaw, P.: Greedy localization, iterative compression, modeled crown reductions: New FPT techniques, and improved algorithm for set splitting, and a novel $2k$ kernelization of vertex cover. In: IWPEC 2004. Lecture Notes in Computer Science, vol. 3162, pp. 127–137 (2004)
9. Downey, R., Fellows, M.: Parameterized Complexity. Springer, New York (1999)

10. Fredman, M., Komlos, J., Szemeredi, E.: Storing a sparse table with $O(1)$ worst case access time. J. ACM **31**, 538–544 (1984)
11. Garey, M., Johnson, D.: Computers and Intractability: A Guide to the Theory of NP-Completeness. Freeman, San Francisco (1979)
12. Hardy, G., Wright, E.: An Introduction to the Theory of Numbers, 5th ed. Oxford University Press, London (1978)
13. Kann, V., Lagergren, J., Panconesi, A.: Approximability of maximum splitting of k-sets and some other APX-complete problems. Inf. Process. Lett. **58**, 105–110 (1996)
14. Kneis, J., Mölle, D., Richter, S., Rossmanith, P.: Divide-and-color. In: WG 2006. Lecture Notes in Computer Science, vol. 4271, pp. 58–67 (2006)
15. Lokshtanov, D., Sloper, C.: Fixed parameter set splitting, linear kernel and improved running time. In: Algorithms and Complexity in Durham 2005. Texts in Algorithmics, vol. 4, pp. 105–113. King's College Press, London (2005)
16. Naor, M., Schulman, L., Srinivasan, A.: Splitters and near-optimal derandomization. In: Proc. 36th IEEE Symp. on Foundations of Computer Science (FOCS 1995), pp. 182–190 (1995)
17. Zhang, H., Ling, C.: An improved learning algorithm for augmented naive Bayes. In: PAKDD 2001. Lecture Notes in Computer Science, vol. 2035, pp. 581–586 (2001)
18. Zwick, U.: Approximation algorithms for constraint satisfaction problems involving at most three variables per constraint. In: Proc. of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 1998), pp. 201–220 (1998)
19. Zwick, U.: Outward rotations: A tool for rounding solutions of semidefinite programming relaxation, with applications to max cut and other problem. In: Proc. of the Thirty-First Annual ACM Symposium on Theory of Computing (STOC 1999), pp. 679–687 (1999)