The Inclusion-Exclusion formula expresses the size of a union of a family of sets in terms of the sizes of intersections of all subfamilies. This paper considers approximating the size of the union when intersection sizes are known for only some of the subfamilies, or when these quantities are given to within some error, or both.

In particular, we consider the case when all \( k \)-wise intersections are given for every \( k \leq K \). It turns out that the answer changes in a significant way around \( K = \sqrt{n} \): if \( K < O(\sqrt{n}) \) then any approximation may err by a factor of \( \Theta(n/K^2) \), while if \( K \geq \Omega(\sqrt{n}) \) it is shown how to approximate the size of the union to within a multiplicative factor of \( 1 - e^{-\Omega(K/\sqrt{n})} \).

When the sizes of all intersections are only given approximately, good bounds are derived on how well the size of the union may be approximated. Several applications for Boolean function are mentioned in conclusion.

1. Introduction

Are all the terms in the inclusion-exclusion formula

\[
|A_1 \cup A_2 \cup \ldots \cup A_n| = \sum_i |A_i| - \sum_{i<j} |A_i \cap A_j| + \sum_{i<j<k} |A_i \cap A_j \cap A_k| - \ldots + (-1)^n |A_1 \cap \ldots \cap A_n|
\]

really necessary? The obvious answer is positive. In the absence of even a single term the size of the union is not uniquely specified. But can the size of the union be approximated well, given only some of the terms? Also, if terms are given to within some error, can the size of the union approximated? The present article answers questions of this general character.

Our interest in these problems arose from some computational considerations: Many computational problems may be viewed as asking for the size of a union of a collection of sets. On some instances it turns out that while computing the size of the union is rather difficult, computing the sizes of members in the family, or even of arbitrary intersections thereof is easy. In these cases, the inclusion-exclusion formula may be used to find the size of the union.

Perhaps the most obvious example is the problem of computing the number of satisfying assignments to a DNF formula (a problem known to be \#P-complete [10]). This problem can be stated as that of computing the size of the union of the sets of assignments that satisfy the various clauses of the DNF formula. The number of
assignments satisfying an intersection of clauses is either zero or $2^m$ where $m$ is the number of variables which appear in none of these clauses. So inclusion-exclusion may be applied to derive the size of the union. This procedure takes time which is exponential in the number of clauses, and seems to be the best algorithm known for this problem when the number of clauses is less than the total number of variables that occur in the formula. In fact, in these cases, this method even seems to be the quickest way known to check whether every assignment satisfies the DNF formula, i.e. to check if the complement, a CNF formula, is satisfiable.

A somewhat more subtle example is Ryser's formula [9] for computing the permanent (also a $\#P$-complete problem [10]). Ryser essentially reduces the problem of computing the permanent to a problem of computing the size of a union of sets, where the sizes of intersections of all subfamilies can be easily computed. The inclusion-exclusion formula is then used to compute the size of the union. This is the quickest method known to compute the permanent, as it requires $2^{n+o(n)}$ operations to compute the permanent of an $n$ by $n$ matrix, instead of the trivial $n!$.

The obvious drawback of using the inclusion-exclusion formula is the fact that it has an exponential number of terms, and that, as mentioned, all terms are necessary, i.e. if the size of the intersection of any subcollection is missing, then the size of the union cannot be computed. This prompted our interest in approximate versions of the inclusion-exclusion formula.

We start, in Section 2, with the following version of this problem: Let $A_1, A_2, \ldots, A_n$ be a collection of sets. Suppose that $\left| \bigcap_{i \in S} A_i \right|$ is given for every subset $S \subseteq [n]$ of cardinality $|S| \leq k$. How well can $\bigcup A_i$ be approximated based only on this information? Equivalently, let $A_1, A_2, \ldots, A_n$ and $B_1, B_2, \ldots, B_n$ be two collections of sets such that for every $S \subseteq [n]$ of cardinality $|S| \leq k$ there holds $\left| \bigcap_{i \in S} A_i \right| = \left| \bigcap_{i \in S} B_i \right|$. How different can $\left| \bigcup A_i \right|$ and $\left| \bigcup B_i \right|$ be?

A naive approach to the problem would be to truncate the exclusion-inclusion formula up to the $k$-terms. This approach is easily seen to fail completely e.g. when all sets are identical.

We give a nearly complete answer to this question, essentially showing that for $k < O(\sqrt{n})$ no good approximation is possible, while for larger $k$ a good approximation is possible. The essence of our main result may be formulated as:

Theorem 1. Let $k$ and $n$ be integers and let $A_1, A_2, \ldots, A_n$ and $B_1, B_2, \ldots, B_n$ be collections of sets where not all $B_i$ are empty and where:

$$\left| \bigcap_{i \in S} A_i \right| = \left| \bigcap_{i \in S} B_i \right|$$

for every subset $S \subseteq [n]$ such that $|S| \leq k$, then

1. For $k \geq \Omega(\sqrt{n})$,

$$\frac{\left| \bigcup_{i=1}^n A_i \right|}{\left| \bigcup_{i=1}^n B_i \right|} = 1 + O\left( e^{-2k/\sqrt{n}} \right)$$

2. For $k \leq O(\sqrt{n})$

$$\frac{\left| \bigcup_{i=1}^n A_i \right|}{\left| \bigcup_{i=1}^n B_i \right|} = O\left( \frac{n}{k^2} \right)$$