Random Sampling and Max-SNP problems

Noga Alon
Schools of Mathematics and Computer Science
Tel Aviv University.
noga@math.tau.ac.il *

Fernandez delaVega ???????

Ravi Kannan
Department of Computer Science,
Yale University.
kannan@cs.yale.edu†

Marek Karpinski ??????

February 22, 2002

Abstract

This is subsection 2.1, to be added to the paper.

1 Introduction

1.1 Notation

2 Existence of Cut Decomposition

In this section, we prove the existence of a certain approximation to any matrix. The approximation will be the sum of a small number of cut-arrays. The sum is taken entry-wise. The proof is elementary and essentially drawn from [6].

Theorem 1 Suppose $A$ is an array on $V_1, V_2, \ldots, V_r$, $N = |V_1| \cdot |V_2| \cdots |V_r|$, and $\epsilon$ is a positive real number. There exist at most $4r/\epsilon^2$ cut arrays whose sum $D$ approximates $A$ well in the sense:

$$
||A - D||_C \leq \epsilon \sqrt{N} ||A||_F \quad (1)
$$

$$
||A - D||_F \leq ||A||_F \quad (2)
$$

The sum of the squares of the coefficients of the cut arrays is at most $4r N||A||_F^2$.

\[ N \]

*supported in part by a USA-Israeli BSF grant, by the Israel Science Foundation and by the Hermann Minkowski Minerva Center for Geometry at Tel Aviv University.

†Supported in part by NSF Grant CCR-9820850.
2.1 Lower Bound on number of Cut Arrays needed

In this subsection we show that the $c(r)/\epsilon^2$ upper estimate for the number of cut arrays in Theorem 1 is tight (up to the dependence on $r$), even if we restrict our attention to $\{-1, 1\}$-arrays $A$, and even if we only require that the sum of the cut arrays $D$ will satisfy (1). Throughout the subsection we assume, whenever this is needed, that $\epsilon$ is sufficiently small as a function of $r$. We also omit all floor and ceiling signs whenever these are not crucial, to simplify the presentation. Note that if we only wish to satisfy (1) in Theorem 1, then its proof implies that $1/\epsilon^2$ cut arrays suffice, as the extra 4’ term appears because of the need to satisfy (3).

The $L_1$-norm of an array $A : V_1 \times V_2 \cdots \times V_r \mapsto R$ is given by

$$||A||_1 = \sum_{(i_1,i_2,\ldots,i_r) \in V_1 \times V_2 \cdots \times V_r} |A(i_1,i_2,\ldots,i_r)|.$$  

The following lemma supplies a lower bound for the cut-norm of an array in terms of its $L_1$-norm. The proof is based on the method of [1].

**Lemma 1** Let $A : V_1 \times V_2 \cdots \times V_r \mapsto R$ be an array. Then its cut norm satisfies

$$||A||_C \geq \frac{||A||_1}{2 \cdot 8^{(r-1)/2} \prod_{j=2}^r |V_j|^{1/2}}.$$  

The proof (following the ideas of [1]) uses a result of Szarek. Let $c_1,c_2,\ldots,c_n$ be a set of $n$ reals, let $\delta_1,\ldots,\delta_n$ be independent, identically distributed random variables, each distributed uniformly on $\{-1,1\}$, and define $X = \sum_i \delta_ic_i$.

**Lemma 2 (Szarek [11])** In the above notation,

$$E(|X|) \geq 2^{-1/2}(c_1^2 + \ldots + c_n^2)^{1/2} (\geq \frac{|c_1| + \ldots + |c_n|}{\sqrt{2n}})$$

**Corollary 1** Let $c_1,\ldots,c_n$ be reals, and let $S$ be a random subset of $\{1, 2, \ldots, n\}$ taken uniformly among all $2^n$ subsets. Let $Y$ be the random variable $Y = \sum_{i \in S} c_i$. Then

$$E(|Y|) = \sum_{S \subseteq \{1,\ldots,n\}} \frac{\sum_{i \in S} c_i}{2^n} \geq \frac{\sum_i |c_i|}{\sqrt{8n}}$$

**Proof:** For every vector $\delta = (\delta_1,\ldots,\delta_n) \in \{-1,1\}^n$ define $S_\delta = \{i : \delta_i = 1\}$ and $S_\delta' = \{i : \delta_i = -1\}$. Then, by the triangle inequality

$$|\sum_{i \in S_\delta} c_i| + |\sum_{i \in S_\delta'} c_i| \geq |\sum_i \delta_ic_i|.$$  

As $\delta$ ranges over all $2^n$ members of $\{-1,1\}^n$, $S_\delta$, as well as $S_\delta'$ range over all $2^n$ subsets of $\{1,2,\ldots,n\}$ implying that $2E(|Y|) \geq E(|X|)$, where $X$ is as above. The result now follows from Lemma 2. \(\Box\)

**Proof of Lemma 1:** We prove, by induction on $t$, that for every $0 \leq t \leq r$ there are subsets $S_{r-t+1} \subset S_{r-t+1} \cdots S_r \subset V_r$ such that

$$\sum_{i_1 \in V_1} \cdots \sum_{i_{r-t} \in V_{r-t}} \sum_{i_{r-t+1} \in S_{r-t+1}} \cdots \sum_{i_r \in S_r} |A(i_1,i_2,\ldots,i_r)| \geq \frac{||A||_1}{8^{t/2} \prod_{j=r-t+1}^r |V_j|^{1/2}}.$$  

(4)
For \( t = 0 \) there is nothing to prove. Assuming the assertion holds for \( t - 1 \leq r \), we prove it for \( t \). For each \((r - t)\)-tuple \( i_1, i_2, \ldots, i_{r-t} \) and each \( i \in V_{r-t+1} \) define

\[
c_i = c_i(i_1, i_2, \ldots, i_{r-t}) = \sum_{i_{r-t+2} \in S_{r-t+2}} \cdots \sum_{i_r \in S_r} A(i_1, i_2, \ldots, i_{t-r-1}, i, i_{t-r+1}, \ldots, i_r),
\]

and apply Corollary 1 with \( n = |V_{r-t+1}| \). Summing the resulting inequalities for all \((i_1, \ldots, i_{r-t}) \in V_1 \times \cdots \times V_{r-t} \) we conclude that the average (over \( S_{r-t+1} \subseteq V_{r-t+1} \)) of the sum

\[
\sum_{i_1 \in V_1} \cdots \sum_{i_{r-t} \in V_{r-t}} \left| \sum_{i_{r-t+1} \in S_{r-t+1}} \cdots \sum_{i_r \in S_r} A(i_1, i_2, \ldots, i_r) \right|
\]

is at least

\[
\frac{1}{\sqrt{8}|V_{r-t+1}|} \frac{||A||_1}{8(t-1)/2 \prod_{j=r-t+2} |V_j|^{1/2}} = \frac{||A||_1}{8(r-1)/2 \prod_{j=2} |V_j|^{1/2}}.
\]

Therefore, there is a set \( S_{r-t+1} \subseteq V_{r-t+1} \) for which (4) holds, showing that it indeed holds for all \( t \leq r \).

In particular, for \( t = r - 1 \) there are sets \( S_2 \subseteq V_2, \ldots, S_r \subseteq V_r \) such that

\[
\sum_{i_1 \in V_1} \sum_{i_2 \in S_2} \cdots \sum_{i_r \in V_r} A(i_1, i_2, \ldots, i_r) \geq \frac{||A||_1}{8(r-1)/2 \prod_{j=2} |V_j|^{1/2}}.
\]

Fixing such sets \( S_i \), either the contribution of the positive terms \( \sum_{i_1 \in S_1} \cdots \sum_{i_r \in V_r} A(i_1, i_2, \ldots, i_r) \) gives at least half of (5), or the contribution of the absolute values of the negative terms gives at least half the sum. In each case we can define \( S_1 \) as the set of those \( i_1 \in V_1 \) that correspond to those contributing terms and conclude that

\[
||A||_C \geq \frac{1}{2} \frac{||A||_1}{8(r-1)/2 \prod_{j=2} |V_j|^{1/2}}.
\]

This completes the proof. \( \square \)

From now on we restrict our attention in this subsection to arrays \( A : V_1 \times V_2 \times \cdots \times V_r \to \{-1, 1\} \) where \( |V_i| = n \) for all \( i \). We need the following simple fact.

**Lemma 3** There exists a family \( F \) of \( r \)-dimensional arrays, each mapping \( V_1 \times V_2 \times \cdots \times V_r \), where \( |V_i| = n \) for each \( i \), into \( \{-1, 1\} \) such that \( |F| \geq 2^{n^2/2} \) and for each two distinct members \( A, B \in F \), \( \|A - B\|_1 > \frac{f_r}{2^n} \).

**Proof:** Let \( H(x) = -x \log_2 x - (1 - x) \log_2(1 - x) \) be the binary entropy function. By the Gilbert-Varshamov bound (see, e.g., [10]), for every (large) \( m \) there are at least \( 2^{(1-H(1/10))m} \geq 2^{m^2/2} \) vectors of length \( m \) over \( \{-1, 1\} \), where the Hamming distance between each pair exceeds \( m/10 \). Taking \( m = n^r \) and viewing these vectors as arrays mapping \( V_1 \times \cdots \times V_r \) to \( \{-1, 1\} \), the desired result follows, as the difference between any two distinct arrays in the collection will have more than \( n^r/10 \) nonzero entries, each of which is either 2 or -2. \( \square \)

We can now prove the main result of this subsection.

**Theorem 2** For every fixed dimension \( r \geq 2 \) there exists some \( c(r) > 0 \) so that for every \( \epsilon > 0 \) there are \( n, N = n^r \) and an \( r \)-dimensional array \( A : V_1 \times \cdots \times V_r \to \{-1, 1\} \), where \( |V_i| = n \) for all \( i \), such that for every array \( D \) which is the sum of less than \( c(r)/\epsilon^2 \) cut arrays,

\[
\|A - D\|_C > \epsilon n^r \quad (= c\sqrt{N}\|A\|_F)
\]
Lemma 1 this implies that for every such $A, B$ follows that the number of arrays $F$ each member of $V \times \cdots \times V_n$ where here we used the fact that $F$ where the last equality follows from the definition of $n$. 

Proof: We prove the theorem for all $\epsilon$ which is sufficiently small as a function of $r$, and with $c(r) = \frac{1}{4r - 40^2 \cdot \epsilon^2 r^2}$. Clearly this implies the result for all $\epsilon$ (with a possibly smaller $c = c(r)$). Define

$$n = \frac{1}{8 \cdot (40)^{2/(r-1)} \epsilon^2 (r-1)^r},$$

and note that $N = n' < 1/(2\epsilon^4)$. By Lemma 3 there is a family $F$ of $2n'/2$ arrays $A : V_1 \times V_2 \times \cdots \times V_r \rightarrow \{-1, 1\}$ such that for every two distinct members $A, B \in F$, $||A - B||_1 > N/5$. By Lemma 1 this implies that for every such $A, B$,

$$||A - B||_c \geq \frac{||A - B||_1}{2 \cdot 8^{(r-1)/2} n^{(r-1)/2}} > \frac{n^{(r+1)/2}}{10 \cdot 8^{(r-1)/2}} = 4n',$$

where the last equality follows from the definition of $n$.

Therefore, $F$ is a large set of arrays, so that the cut-distance between any pair of them is large. To complete the proof we show that at least one member of $F$ cannot be approximated well (in the cut metric) by a sum of a small number of cut arrays. To do so, suppose that for each member $A$ of $F$ there is an array $D$ which is a sum of at most $t$ cut arrays, such that $||A - D||_c \leq c n'$. Call a cut-array $c$-nice if it is an array of the form $CUT(S_1, S_2, \ldots, S_i; d)$ where $d$ is an integral multiple of $\epsilon/t$. An obvious rounding procedure implies that for each member of $F$ there is an array $D$ which is the sum of at most $t$ $c$-nice cut arrays, such that $||A - D||_c \leq 2en'$.

We next prove an upper bound for the total possible number of such arrays $D$. Note, first, that as $n' < 1/(2\epsilon^4)$, the absolute value of no entry of such a $D$ can exceed $1 + 1/\epsilon^3 < 2/\epsilon^3$ (since otherwise the cut-norm of $A - D$ would exceed $2en'$ simply by considering a single entry). As each entry of $D$ is also an integral multiple of $\epsilon/t$ it follows that there are at most $4t/\epsilon^4$ possibilities for each such entry. There are at most $2^{nrt}$ possibilities for choosing the sets $S_1, \ldots, S_r$ in each cut array $CUT(S_1, \ldots, S_i; d)$, and as $D$ is the sum of $t$ such arrays there are at most $2^{nrt}$ possibilities for choosing the defining sets of all of them. Once these are chosen, we have to choose the densities $d$ of these arrays. Each of those is an integral multiple of $\epsilon/t$, but the trouble is that its absolute value may be large (as there may be cancellations between them, while forming $D$). It is thus better to bound the number of possibilities of all these densities as follows. Let $d_1, \ldots, d_t$ be the densities. Since we have already chosen all sets $S_i$ in all the cut arrays whose sum is $D$, we can express each entry of $D$ as a sum of a subset of the densities $d_i$. At most $t$ of the characteristic vectors of these subsets span all the characteristic vectors of all other subsets we have, and thus if we are given the values of $D$ in these entries, we can solve for all other entries of $D$. There are at most $n^rt$ ways to choose $t$ entries of $D$, and then there are at most $(4t/\epsilon^4)^t$ possibilities for the values of $D$ in these entries (as each entry is an integral multiple of $\epsilon/t$ whose absolute value does not exceed $2/\epsilon^3$.) Therefore, the total number of possible arrays $D$ is at most

$$n^{rt} \left( \frac{4t}{\epsilon} \right)^t 2^{nrt}.$$

Each member of $F$ is within cut-distance smaller than $2en'$ from at least one of these arrays $D$, and the cut-distance between any two distinct members of $F$ exceeds $4en'$, by (6). It thus follows that the number of arrays $D$ is at least as large as $F$, implying that

$$\log |F| = \frac{n'}{2} \leq rt \log n + t \log(4t/\epsilon) + nrt < 2trn,$$

where here we used the fact that $n$ is much bigger than $\log n + \log(4t/\epsilon)$. The last inequality implies that

$$t \geq \frac{n^{r-1}}{4r} = \frac{1}{4r \cdot 40^2 \cdot 8^{r-1}\epsilon^2},$$

completing the proof. $\square$
References


