ON SPARSE APPROXIMATIONS TO RANDOMIZED STRATEGIES AND CONVEX COMBINATIONS

INGO ALTHÖFER

Fakultät für Mathematik
Universität Bielefeld
Universitätsstraße
33 615 Bielefeld
Germany

ABSTRACT. A randomized strategy or a convex combination may be represented by a probability vector $p = (p_1, \ldots, p_m)$. p is called sparse if it has only few positive entries. This paper presents an Approximation Lemma and applies it to matrix games, linear programming, computer chess, and uniform sampling spaces. In all cases arbitrary probability vectors can be substituted by sparse ones (with only logarithmically many positive entries) without loosing too much performance.

Short running title: Sparse Approximations

1. Introduction

We present the following

Approximation Lemma:

Let $A = (a_{ij})$ be an $m \times n$–matrix over the real numbers with $0 \leq a_{ij} \leq 1$ for $1 \leq i \leq m$, $1 \leq j \leq n$. Let $p = (p_1, \ldots, p_m)$ be a probability vector, i.e., $0 \leq p_i$ for all i and $\sum_{i=1}^{m} p_i = 1$, and $\varepsilon > 0$ any positive constant. Then there exists another probability vector $q = (q_1, \ldots, q_m)$ with at most $k = \left\lceil \frac{\log 2n}{2\varepsilon^2} \right\rceil$ many positive coordinates q_i such that

$$\left| \sum_{i=1}^{m} p_i a_{ij} - \sum_{i=1}^{m} q_i a_{ij} \right| \leq \varepsilon \qquad \text{for all} \qquad j = 1, \ldots, n \ .$$

More precisely, the probability vector q can be chosen such that $q_i = \frac{k_i}{k}$ with natural numbers k_i for all $i = 1, \ldots, m$. q is called an ε*–approximation* to p .

Here $\log(.)$ means ‘logarithmus naturalis’, and $\lceil x \rceil = \min\{k \in \mathbb{Z} \mid k \geq x\}$ is the function that rounds to the next larger integer.

Observe that the bound k in the lemma depends only on the number n of columns of A and on the approximation level ε , but not on the number m of rows, the entries a_{ij} of A , and the probabilities p_i .

The paper is organized as follows: In Section 2 we illustrate some applications of the Approximation Lemma to matrix games, linear programming, computer chess, and uniform sampling spaces. Section 3 contains the proof of the lemma. In Section 4 we have a look at algorithmic aspects of constructing the sparse probability distribution q . In Section 5 we show that the bound $k = \left\lceil \frac{\log 2n}{2\varepsilon^2} \right\rceil$ is asymptotically optimal up to the multiplicative constant $4 \log 2$. Section 6 contains a discussion of miscellaneous ideas. Especially, in Paragraph 6 C) we formulate a slightly more general version of the Approximation Lemma.

2. Applications

A) Matrix Games

In the theory of zero–sum matrix games the following Theorem holds (see for instance the books [Kar 59] and [Ros 81] as references).

Theorem. *For every* $m \times n$*–matrix* A *with real valued entries there is some* $v \in \mathbb{R}$ *and two probability distributions* $p = (p_1 \dots p_m)$ *and* $r = (r_1 \dots r_n)$ *such that*

$$\text{(i)} \qquad \sum_{i=1}^{m} p_i a_{ij} \geq v \qquad \textit{for all } j = 1 \dots n.$$

$$\text{(ii)} \qquad \sum_{j=1}^{n} a_{ij} r_j \leq v \qquad \textit{for all } i = 1 \dots m\ .$$

$v = v(A)$ is called the game theoretic value or minimax value of A . p and r are optimal randomized strategies for Max and Min , respectively. p and r need not be unique, whereas the value $v(A)$ is unique. In case of $m \geq n$ Max can choose his optimal strategy p in such a way that at most n many of the p_i are positive [Kar 59, Theorem 2.4.3 on page 50].

A strategy $q = (q_1 \dots q_m)$ for Max is called ε*–approximation* for A if

$$\sum_{i=1}^{m} q_i a_{ij} \geq v(A) - \varepsilon \qquad \text{for all} \qquad j = 1 \dots n\ . \tag{2.1}$$

The Approximation Lemma (called AL in the sequel) guarantees the existence of ε–approximations with at most $\left\lceil \frac{\log 2n}{2\varepsilon^2} \right\rceil$ many positive q_i if the entries of A are all in the interval $[0,1]$. From the proof of the AL in Section 3 it will follow that even $\left\lceil \frac{\log n}{2\varepsilon^2} \right\rceil$ positive q_i will suffice for ε–approximations . This stems from the one–sidedness of the inequalities (2.1) ($\sum q_i a_{ij} \leq v(A) + \varepsilon$ is not requested). From Section 5 it will follow that there are matrices A where every ε–approximation for Max requires $\Omega\left(\frac{\log n}{\varepsilon^2}\right)$ many positive q_i .

Similar approximation results with only "logarithmic randomization" also hold for Bi–matrix games where the payoffs for the 2 players are fixed by two different matrices A and B .

B) Linear Programming

Consider the following special type of a linear program.

Maximize $\sum_{i=1}^{m} c_i x_i$ under the constraints

$$
\begin{aligned}
& x_i \geq 0 \quad \text{for all } i, \\
& \sum_{i=1}^{m} x_i \leq 1, \\
& a_{11}x_1 \quad + \cdots + \quad a_{m1}x_m \leq b_1, \\
& \qquad\qquad \vdots \\
& a_{1n}x_1 \quad + \cdots + \quad a_{mn}x_m \leq b_n \ .
\end{aligned}
$$

Assume that $0 \leq c_i \leq 1$ for all i, $0 \leq a_{ij} \leq 1$ for all i, j, and $\frac{1}{2} \leq b_j \leq 1$ for all j. Let $(y_1 \ldots y_m)$ be an optimal solution with $C := \sum_{i=1}^{m} c_i y_i \leq 1$, and $\varepsilon > 0$ a constant.

Then there is some other admissible vector $(z_1 \ldots z_m)$ such that

(i) $\sum_{i=1}^{m} c_i z_i \geq C - 3\varepsilon$

and

(ii) at most $\left\lceil \frac{\log 2(n+1)}{2\varepsilon^2} \right\rceil$ many of the z_i are not equal to zero.

Proof: Let $y_0 = 1 - \sum_{i=1}^{m} y_i$. $(y_0, y_1 \ldots y_m)$ is a probability distribution. By the AL there is another probability distribution $z' = (z'_o, z'_1 \ldots z'_m)$ such that

(a) $0 \cdot z'_0 + a_{1j} z'_1 + \cdots + a_{mj} z'_m \leq b_j + \varepsilon$ for all $j = 1 \ldots n$,

(b) $\sum_{i=1}^{m} c_i z'_i \geq C - \varepsilon$

and

(c) at most $\left\lceil \frac{\log 2(n+1)}{2\varepsilon^2} \right\rceil$ many of the z'_i are positive.

Now modify z' to z by setting $z_i = \frac{1}{1+2\varepsilon} \cdot z'_i$ for $i = 1 \ldots m$.
From (a) we get for all j

$$\sum_{i=1}^{m} a_{ij} z_i = \frac{1}{1+2\varepsilon} \sum_{i=1}^{m} a_{ij} z'_i \leq \frac{1}{1+2\varepsilon}(b_j + \varepsilon) \leq (1 - 2\varepsilon) b_j + \varepsilon = b_j + \varepsilon(1 - 2b_j) \leq b_j \ ,$$

as $b_j \geq \frac{1}{2}$.

From (b) we get

$$\sum_{i=1}^{m} c_i z_i = \frac{1}{1+2\varepsilon} \sum_{i=1}^{m} c_i z_i' \geq \frac{1}{1+2\varepsilon}(C-\varepsilon) \geq (1-2\varepsilon)C - \varepsilon = C - \varepsilon(1+2C) \geq C - 3\varepsilon \ .$$

If instead of $b_j \geq \frac{1}{2}$ we demand $b_j \geq \delta > 0$, part (i) of our result is weakened to $\sum_{i=1}^{n} c_i z_i \geq C - \left(1 + \frac{1}{\delta}\right) \varepsilon$.

The following example shows that these "strange" lower bounds on the b_j are necessary to guarantee ε–approximations with only few positive z_i : Consider inequalities $x_i \leq \frac{1}{m}$ for $i = 1, \ldots, m$ and $c_i = 1$ for all i . Then $\left(\frac{1}{m}, \ldots, \frac{1}{m}\right)$ is an optimal solution with $C = 1$. But any feasible point z with only k positive z_i has $\sum_{i=1}^{m} c_i z_i \leq \frac{k}{m}$.

C) Linear Evaluation Functions in Computer Chess and Other Games

A standard technique in computer game playing is game tree search. The core of the basic Shannon–A–strategy [Sha 50] is as follows: The game is represented by a rooted tree. The nodes of the tree are legal positions, the arcs of the tree stand for the moves. The root is the current position for which a move decision has to be made.

In real world games like chess or *Go* the whole trees are terribly large. So in computer game playing only a rather "small" part (the *search tree*) of the tree is built up; and the strengths of the artificial leaves of this search tree are estimated by evaluation functions based on heuristic values. Often such an evaluation function f is a convex combination of elementary evaluation functions $f_1 \ldots f_m$, i.e. $f = \sum_{i=1}^{m} p_i f_i$ for some probability vector $(p_1 \ldots p_m)$. Typically the number m is rather large.

Now assume that there is some "perfect" evaluation function $f = \sum_{i=1}^{m} p_i f_i$, for instance for the game of chess, with the following properties

(i)	$0 \leq f_i(x) \leq 1$	for every chess position x and every f_i ,
(ii)	$\frac{6}{7} < f(x) \leq 1$	for every position x that is a game–theoretic win for player White,
	$\frac{3}{7} < f(x) < \frac{4}{7}$	for every x whose game–theoretic value is a draw,
	$0 \leq f(x) < \frac{1}{7}$	for every x that is a game–theoretic loss for White.

(It is currently unknown for what classes $\{f_1 \ldots f_m\}$ of elementary functions such a convex combination exists.)

In the game of chess there are less than 10^{53} legal positions. So we may apply the AL with $n = 10^{53}$ and $\varepsilon = \frac{1}{7}$, and get:

If an f like above exists there exists also another probability distribution $q = (q_1 \dots q_m)$ such that q and the convex combination $g = \sum q_i f_i$ satisfy

(i) at most $\left\lceil \frac{\log 2\cdot 10^{53}}{2\cdot\left(\frac{1}{7}\right)^2} \right\rceil = 3007$ many of the coefficients q_i are not equal to zero,

(ii) $\frac{5}{7} < g(x) \leq 1$ for every x that is a game–theoretic win for White,

$\frac{2}{7} < g(x) < \frac{5}{7}$ for every drawing position x,

$0 \leq g(x) < \frac{2}{7}$ for every x that is a game–theoretic loss for White.

So the "simple" function g (with at most 3007 terms) is also a perfect separator of the 3 classes White–wins, draws, and White–losses.

In the game of *Go* on the 19×19–board there are less than 3^{361} legal positions. As there are no draws in *Go* (by the komi rule) we may assume a perfect evaluation function $f = \sum p_i f_i$ with the properties

(i) $0 \leq f_i(x) \leq 1$ for every position x and every f_i,

(ii) $\frac{3}{4} < f(x) \leq 1$ for every position x that is a game–theoretic win for Black,

$0 \leq f(x) < \frac{1}{4}$ for every x that is a game–theoretic loss for Black.

Applying the AL with $n = 3^{361}$ and $\varepsilon = \frac{1}{4}$ we see that the existence of an f like above implies the existence of another convex combination $g = \sum_{i=1}^{m} q_i f_i$ such that

(i) at most $\left\lceil \frac{\log 2\cdot 3^{361}}{2\cdot\left(\frac{1}{4}\right)^2} \right\rceil = 3179$ many of the coefficients q_i are not equal to zero,

(ii) $\frac{1}{2} < gx) \leq 1$ for every Black–win x,

$0 \leq g(x) < \frac{1}{2}$ for every Black–loss x.

Of course similar statements can be made for many other games, even for games with a random element like Backgammon.

In the computer chess family there is some permanent discussion about "knowledge versus search" [CT 84]: Should one use more complicated evaluation functions and smaller search trees or vice versa? Our result may be seen as a hint that very complicated convex combinations of elementary functions will not be the king's way to optimal computer chess play [Alt 93].

D) Uniform Sampling Spaces

Consider a finite set $C = \{1, \ldots, m\}$ and functions $f_i : C \to [0,1]$ for $i = 1, \ldots, n$. Let $c_i = \frac{1}{m} \sum_{j=1}^{m} f_i(j)$ denote the average level of f_i. Given some approximation level $\varepsilon > 0$, we are interested in small subsets $B \subset C$ such that the average levels

$$b_i = \frac{1}{|B|} \sum_{j \in B} f_i(j)$$

of f_i on B satisfy $| c_i - b_i | \leq \varepsilon$ for all i. B is called an ε–uniform sampling space for $\{f_1 \ldots f_n\}$.

From the AL we see that there exist ε–uniform sampling spaces with at most $k = \left\lceil \frac{\log 2n}{2\varepsilon^2} \right\rceil$ many elements. Observe that this k depends on n and ε, but not on the size m of the set C.

Example:

Let $C = \{+1, -1\}^t$ be the t–dimensional hypercube. We look at bilinear functions

$$f_i'(x_1 \ldots, x_t) = \sum_{r=1}^{t} d_r x_r + \sum_{r=1}^{t-1} \sum_{s=r+1}^{t} d_{rs} x_r x_s$$

with $d_r \in \{0,1\}$, $d_{rs} \in \{0,1\}$ for all r, s and renormalize them to $f_i = \frac{1}{e_i} f_i' + 1$ by appropriate constants $e_i \in \mathbb{R}_+$ such that $0 \leq f_i(x) \leq 1$ for all $x \in C$. There are $n = 2^t \cdot 2^{\binom{t}{2}} = 2^{\frac{t(t+1)}{2}}$ many different choices for the d_r and d_{rs}. $\log 2^{\frac{t(t+1)}{2}} = \frac{t(t+1)}{2} \cdot \log 2$, hence this class of functions has ε–uniform sampling spaces with only quadratically many elements in t, whereas the set C has exponentially many elements in t. Analogous results hold for tri– and higher multilinear forms.

3. The Proof of the Approximation Lemma

The existence of the probability vector q in the lemma is proved by a random choice argument [ES 74].

The matrix A has m rows $1, 2, \ldots, m$. We define k independent and identically distributed random variables $X_1, \ldots, X_k$ with values in $\{1, \ldots, m\}$ and $\text{Prob}\{X_1 = i\} = p_i$ for $i = 1, \ldots, m$. The probabilities $q_i = \frac{k_i}{k}$ will result from the X_t by setting

$k_i = \#\{t | 1 \leq t \leq k \text{ and } X_t = i\}$.

In our proof we will use an exponential upper bound for the probability of large deviations in sums of independent and bounded random variables.

Hoeffding's Inequality [H 63]:

Let $Y_1, \ldots, Y_k$ be independent random variables, with $0 \le Y_t \le 1$ for each t, let $\overline{Y} = \frac{1}{k} \sum_{t=1}^{k} Y_t$ and let $E\overline{Y}$ be its expectation. Then for all $\varepsilon > 0$

(i) $\mathrm{Prob}\{\overline{Y} - E\overline{Y} \ge \varepsilon\} \le \exp\{-2\varepsilon^2 k\}$, and

(ii) $\mathrm{Prob}\{\overline{Y} - E\overline{Y} \le -\varepsilon\} \le \exp\{-2\varepsilon^2 k\}$.

See also the survey paper of McDiarmid [McD 89] for a nice presentation of this inequality and many applications.

Now fix some column j of A. Setting $Y_t = a_{X_t j}$, the variables Y_t satisfy $0 \le Y_t \le 1$ and $EY_t = \sum_{i=1}^{m} p_i a_{ij}$ for all t. Hence, we can apply Hoeffding's Inequality to get an upper bound for the probability that the deviation for column j is larger then ε:

$$\mathrm{Prob}\left\{ \left| \sum_{i=1}^{m} p_i a_{ij} - \frac{1}{k} \sum_{t=1}^{k} a_{X_t j} \right| > \varepsilon \right\} < 2\exp(-2\varepsilon^2 k) .$$

We can write " $>$ " and " $<$ " instead of " $\ge$ " and " $\le$ ", respectively, as the X_t have only finite ranges.

Combining these results for all n columns, we get

$$\left.\begin{aligned} &\mathrm{Prob}\left\{ \left| \sum_{i=1}^{m} p_i a_{ij} - \frac{1}{k} \sum_{t=1}^{k} a_{X_t j} \right| > \varepsilon \text{ for at least one column } j \right\} \\ &\le \sum_{j=1}^{n} \mathrm{Prob}\left\{ \left| \sum_{i=1}^{m} p_i a_{ij} - \frac{1}{k} \sum_{t=1}^{k} a_{X_t j} \right| > \varepsilon \right\} \\ &< 2n\exp(-2\varepsilon^2 k) . \end{aligned}\right\} \tag{3.1}$$

Choosing $k \ge \frac{\log 2n}{2\varepsilon^2}$, this last term becomes smaller than or equal to 1. Hence the random construction gives deviation $\le \varepsilon$ for all columns with positive probability, hence there must be at least one realization of $(X_1, \ldots, X_k)$ with the desired property.

■

4. Algorithmic Aspects of the Approximation Lemma

The Approximation Lemma is a result of existence. The proof only says: "If you choose k rows independently according to the distribution p then the resulting distribution q (with $q_i = \frac{1}{k}\cdot$(# row i in the k–sample)) is an ε–approximation with probability > 0 ."

There is one crude way to turn this statement into a fast probabilistic construction by allowing slightly more than $k = \left\lceil \frac{\log 2n}{2\varepsilon^2} \right\rceil$ many independent choices, namely $k' = \left\lceil \frac{\log 4n}{2\varepsilon^2} \right\rceil$ many. The resulting distribution q' (with at most k' many positive q'_i) will be an ε–approximation to p with probability $> \frac{1}{2}$ (see (3.1) in Section 3). If this first trial of k' row choices was not successful, make an independent second trial and so on. In the average less than $1 + \frac{1}{2} + \frac{1}{4} + \cdots = 2$ trials of k' choices each will be needed.

However, in practice a strategy with a sequential stopping rule and several trials will typically give better results:

> Pick the independent rows one by one, and after each step check whether the resulting distribution is an ε–approximation . Stop if yes, otherwise continue. If after k' steps the distribution q' is still not okay, stop.

Repeat this procedure several times. With very high probability ($> 1 - 2^{-t}$ for t trials) there will have been at least one successful trial. From all successful trials choose the shortest or the "best" one. For many distributions p and matrices A the bound k from our proof will not be tight. Especially in these cases a sequential strategy like the one described above will need less time and produce better results in the average.

In [KY 76] the complexity of random number generation is investigated. A random choice according to a distribution $(p_1, \ldots, p_m)$ costs approximately $H(p)$ fair coin flips in the average. Here $H(p) = -\sum_{i=1}^{m} p_i \log_2 p_i$ is the information theoretic entropy function (log with base 2 !). So one trial to generate the ε–approximation q will cost about $k' \cdot H(p)$ fair coin flips, whereas the resulting distribution q will have entropy $H(q) \leq \log k' \approx \log\log n$ only.

An interesting task will be to substitute the *probabilistic* constructions above by a fast *deterministic* one.

Open Problem: Are there fast deterministic algorithms for the generation of ε–approximations q with at most $\left\lceil \frac{\log 2n}{2\varepsilon^2} \right\rceil$ many positive q_i ?

Maybe, this problem can be solved by using the method of conditional probabilities, as described in [ASE 92].

5. The Converse

In this section we show that the upper bound $\frac{\log 2n}{2\varepsilon^2}$ in the Approximation Lemma is asymptotically (for large n) optimal up to the multiplicative constant $4\log 2 \approx 2.77$. This is done by using matrices with entries from the two–element set $\{0,1\}$ only. Let us start with a simple example.

Let m be some even integer and $d = \frac{m}{2}$. We investigate a matrix A with m rows and $n = \binom{m}{d}$ columns. The columns are the collection of all $\{0,1\}$–vectors of length m with exactly d '1'– and d '0'–entries. Table 5.1 shows such a matrix for $m = 4$. $\binom{m}{d} \geq \frac{1}{m+1} \cdot 2^m$, hence $\log n \geq m \log 2 - \log(m+1) \approx m \log 2$.

1	1	1	0	0	0
1	0	0	1	1	0
0	1	0	1	0	1
0	0	1	0	1	1

Table 5.1

Now assume some real valued vector $(p_1 \dots p_m)$ with $\sum_{i=1}^{m} p_i = 1$ and at most d many positive p_i . Then there is some column $(c_1 \dots c_m)$ in A such that $\sum_{i=1}^{m} p_i c_i = 1$. Hence every ε–approximation $(q_1 \dots q_m)$ to the uniform distribution $\left(\frac{1}{m} \dots \frac{1}{m}\right)$ requires at least $\frac{m}{2} + 1$ many positive q_i if $\varepsilon < \frac{1}{2}$. $\frac{\log 2n}{2 \cdot \left(\frac{1}{2}\right)^2} \approx (2\log 2) \cdot m = (4 \log 2) \cdot \frac{m}{2}$.

The general converse for fixed ε in $\left(0, \frac{1}{2}\right)$ is formulated in

Theorem 5.1. *Consider an even integer* $m \geq 2$ *and some constant* $0 < \varepsilon < \frac{1}{2}$ *for which* $\left(\frac{1}{2} + \varepsilon\right) \frac{m}{2} \in \mathbb{N}$. *Then there exists a matrix* A *with* m *rows and at most* $n = 2^{4\varepsilon^2 m} \cdot \frac{m^4}{4} \log m + 1$ *columns with* $\frac{m}{2}$ *'0'– and* $\frac{m}{2}$ *'1'–entries in each column such that for every real valued vector* $(p_1 \dots p_m)$ *with* $\sum_{i=1}^{m} p_i = 1$ *and at most* $\frac{m}{2}$ *many* $p_i \neq 0$ *there is some column* $(c_1 \dots c_m)$ *in* A *for which* $\sum_{i=1}^{m} p_i c_i \geq \left(\frac{1}{2} + \varepsilon\right)$.

Hence especially every δ–approximation to the uniform distribution $\left(\frac{1}{m} \dots \frac{1}{m}\right)$ requires at least $\frac{m}{2} + 1$ many positive entries if $\delta < \varepsilon$. Observe that not only probability distributions $(p_1, \dots, p_m)$ are covered by this result.

Forgetting the $(+1)$ term in the expression for n and later the $\log m$ terms we have

$$\begin{aligned} \frac{\log 2n}{2\varepsilon^2} &\approx \frac{4\log m + \log\log m - \log 2 + 4\varepsilon^2 m \log 2}{2\varepsilon^2} \\ &\approx (2\log 2) \cdot m \ . \end{aligned}$$

Hence the Approximation Lemma is optimal up to the multiplicative factor $4 \cdot \log 2$.

The proof of Theorem 5.1 is rather complicated. All its details may be found in the Technical Report [Alt 93a]. In this paper we sketch the key ideas only. We start with some notation. Let $M = \{1 \dots m\}$ be the set of rows, let S be the set of all subsets of M with $d := \frac{m}{2}$ elements, and let X be the set of all ordered sequences $(x_1 \dots x_d)$ such that $\{x_1 \dots x_d\} \in S$.

Next we show how the continuous problem (Theorem 5.1 makes a statement for a continuous set of probability distributions) can be reduced to a discrete one.

Lemma 5.2. *Let* $(c_1 \dots c_d) \in \{0,1\}^d$ *with* $\sum_{i=1}^{k} c_i \geq \left(\frac{1}{2} + \varepsilon\right) k$ *for all* $k = 1, \dots, d-1$ *, and* $\sum_{i=1}^{d} c_i = \left(\frac{1}{2} + \varepsilon\right) d$ *. Then* $\sum_{i=1}^{d} p_i c_i \geq \frac{1}{2} + \varepsilon$ *for all real–valued vectors* $(p_1 \dots p_d)$ *with* $p_1 \geq p_2 \geq \dots \geq p_d$ *and* $\sum_{i=1}^{d} p_i = 1$ *.*

Thus, in order to prove the theorem, we only need the following statement: For every ordered sequence $x = (x_1 \dots x_d) \in X$ there is some column $(c_1 \dots c_m)$ in A such that

$$\sum_{i=1}^{k} c_{x_i} \geq \left(\frac{1}{2} + \varepsilon\right) k \quad \text{for all} \quad k = 1, \dots, d-1, \quad \text{and} \quad \sum_{i=1}^{d} c_{x_i} = \left(\frac{1}{2} + \varepsilon\right) d \ .$$

The existence of a matrix A with this property will be guaranteed by an application of the

Covering Lemma [Ahl 79]:

Let $H = (X, \mathcal{E})$ be a finite hypergraph with $\min_{x \in X} \deg x = r$. Then there exists a covering of X by a set $C \subset \mathcal{E}$ for hyperedges where

$$|C| \leq \frac{|\mathcal{E}|}{r} \log |X| + 1 \ .$$

Here a hypergraph is given by a set X of vertices and a set $\mathcal{E} = \{E_1 \dots E_t\}$ of subsets of X , the hyperedges. For $x \in X$ we write $\deg x = |\{E_i \in \mathcal{E} | x \in E_i\}|$.

In our application X is the set of ordered sequences $x = (x_1 \dots x_d)$ and $\mathcal{E}$ is the set of columns $(c_1 \dots c_m) \in \{0,1\}^m$ with $\sum_{i=1}^{m} c_i = d$. A vertex $x = (x_1 \dots x_d)$ is called an element of the hyperedge $c = (c_1 \dots c_m)$ if

$$\sum_{i=1}^{k} c_{x_i} \geq \left(\frac{1}{2}+\varepsilon\right) k \quad \text{for all} \quad k = 1, \ldots, d-1, \quad \text{and} \quad \sum_{i=1}^{d} c_{x_i} = \left(\frac{1}{2}+\varepsilon\right) d \ .$$

$$|X| = m \cdot (m-1) \cdot \ldots \cdot (m-d+1) \leq m^d \ . \tag{5.1}$$

Now the auxiliary set S comes into play. We say that a hyperedge $(c_1 \ldots c_m)$ *touches* a set $Z = \{z_1 \ldots z_d\} \in S$ if $\sum_{i=1}^{d} c_{z_i} = \left(\frac{1}{2}+\varepsilon\right) d$. For $Z \in S$ we define $X(Z) = \left\{(x_1 \ldots x_d) \in X \mid \{x_1 \ldots x_d\} = Z\right\}$. $\mid X(Z) \mid = d!$.

Lemma 5.3. *Let* $c = (c_1 \ldots c_m) \in \mathcal{E}$ *be a hyperedge and* $Z \in S$ *a set that is touched by* c *. Then*

$$\mid \{x \in X(Z) | x \in c\} \mid \geq (d-1)!$$

Lemma 5.4. *Every hyperedge* c *contains at least* $\binom{d}{(\frac{1}{2}+\varepsilon)d} \cdot \binom{d}{(\frac{1}{2}-\varepsilon)d} \cdot (d-1)!$ *many vertices* $x \in X$ *.*

By the symmetry of our hypergraph $(X, \mathcal{E})$ all hyperedges have the same number $|c|$ of vertices, and all vertices have the same degre r . Thus

$$r \cdot |X| = |\mathcal{E}| \cdot |c|, \qquad \text{and this is equivalent to}$$

$$r = \frac{|\mathcal{E}| \cdot |c|}{|X|} \ . \tag{5.2}$$

Before applying the Covering Lemma we restate some known bounds for binomial coefficients.

Lemma 5.5. *Let* d *be some natural number and* $0 \leq \varepsilon \leq \frac{1}{2}$ *a constant such that* $\left(\frac{1}{2}+\varepsilon\right) d \in \mathbb{N}$ *. Then*

$$\frac{1}{d+1} \cdot 2^{(1-4\varepsilon^2)d} \leq \binom{d}{\left(\frac{1}{2}+\varepsilon\right) d} \leq 2^{\left(1-\frac{2}{\log 2}\varepsilon^2\right)d} \ .$$

Combining the Covering Lemma, equation (5.2), Lemma (5.4), Lemma (5.5), and inequality (5.1), we see that there exists a covering C of X of size

$$
\begin{aligned}
|C| &\le \frac{|\mathcal{E}|}{r}\log|X|+1 \\
&= \frac{|X|}{|c|}\log|X|+1 \\
&\le \frac{|X|}{\binom{d}{(\frac{1}{2}+\varepsilon)d}\cdot\binom{d}{(\frac{1}{2}-\varepsilon)d}\cdot(d-1)!}\log|X|+1 \\
&\le \frac{m\cdot(m-1)\cdot\ldots\cdot(m-d+1)}{(d-1)!}\cdot\frac{(d+1)^2}{2^{2(1-4\varepsilon^2)d}}\log|X|+1 \\
&= \binom{m}{d}\cdot d\cdot\frac{(d+1)^2}{2^{2(1-4\varepsilon^2)d}}\log|X|+1 \\
&\le 2^m\cdot\frac{d(d+1)^2}{2^{2(1-4\varepsilon^2)d}}\cdot d\log m+1 \\
&= d^2(d+1)^2\log m\cdot 2^{8\varepsilon^2 d}+1\ . \qquad (5.3)
\end{aligned}
$$

For $m \ge 2$ we have $d+1=\frac{m}{2}+1\le m$, hence (5.5) $\le \frac{m^4}{4}\log m\cdot 2^{4\varepsilon^2 m}+1$. This completes the proof of Theorem 5.1.

For small $\varepsilon>0$ we have $\binom{d}{(\frac{1}{2}+\varepsilon)d}\approx 2^{\left(1-\frac{2}{\log 2}\varepsilon^2\right)d}$. Hence in this case there exists a covering C with $n=|C|\approx\frac{m^4}{4}\log m\cdot 2^{\frac{2}{\log 2}\varepsilon^2 m}$, thus

$$
\frac{\log 2n}{2\varepsilon^2}\approx\frac{\frac{2}{\log 2}\varepsilon^2 m\log 2+\log 2}{2\varepsilon^2}\approx m
$$

which shows that for small ε the Approximation Lemma is asymptotically optimal up to a multiplicative factor 2 .

6. Miscellaneous Ideas

A) For $\varepsilon>0$ fixed and n large we have $n\gg\frac{\log 2n}{2\varepsilon^2}$. Hence the Approximation Lemma may save many rows in this case, for instance if $m\approx n$. For small n the term $\frac{1}{\varepsilon^2}$ may dominate the expression. Table 6.1 exhibits for some values of ε the smallest $n=n^*(\varepsilon)$ for which $n>\left\lceil\frac{\log 2n}{2\varepsilon^2}\right\rceil$. In general $n^*(\varepsilon)>\frac{\log(\frac{1}{\varepsilon})}{\varepsilon^2}$ for all $\varepsilon>0$, and $\lim_{\varepsilon\to 0} n^*(\varepsilon)\cdot\frac{\varepsilon^2}{\log\left(\frac{1}{\varepsilon}\right)}=1$.

ε	$n^*(\varepsilon)$
0.5	6
0.3	23
0.1	325
0.03	$5,133$
0.01	$58,337$
0.001	$8,313,256$

Table 6.1

B) Given a cost vector $(c_1 \dots c_m) \in \mathbb{R}^m$, one can define the average cost of a probability distribution $p = (p_1 \dots p_m)$ by $c(p) = \sum_{i=1}^{m} c_i p_i$. Sometimes we may be interested in sparse ε–approximations q with small average cost $c(q)$.

Theorem 6.1. *Consider the situation of the AL and additionally a cost vector $(c_1 \dots c_m)$ with $c_i \geq 0$ for all i . Then there exists an ε–approximation q to p with the following properties:*

(i) at most $\left\lceil \frac{\log 4n}{2\varepsilon^2} \right\rceil$ many q_i are not equal to zero,

(ii) $c(q) < 2c(p)$.

Proof:

Make the random construction of Section 3 (and/or Section 4) with $\left\lceil \frac{\log 4n}{2\varepsilon^2} \right\rceil$ many independent choices. With probability $> \frac{1}{2}$ this will give an ε–approximation q to p . On the other hand the expected cost of q is $c(p)$, hence by Markow's inequality $\text{Prob}\{c(q) \geq 2c(p)\} \leq \frac{1}{2}$, as the cost vector is non–negative. Putting both results together guarantees the existence of a distribution q with the desired properties.

$\blacksquare$

C) Here we formulate a slightly more general version of the Approximation Lemma.

For a real valued $m \times n$–matrix A let us denote the maximum difference within a column by

$$\text{diff}(A) = \max_{1 \leq j \leq n} \left\{ \max_{1 \leq i \leq k \leq m} |a_{ij} - a_{kj}| \right\} .$$

Then the Approximation Lemma may be restated as follows:

"Let A be a real valued $m \times n$–matrix , $(p_1, \dots, p_m)$ a probability distribution, and $\varepsilon > 0$ any positive constant. Then there exists another probability vector $(q_1, \dots, q_m)$ with at most $k = \left\lceil \frac{\log 2n}{2\varepsilon^2} \right\rceil$ many positive coordinates q_i such that

$$\left| \sum_{i=1}^{m} p_i a_{ij} - \sum_{i=1}^{m} q_i a_{ij} \right| \leq \varepsilon \cdot \text{diff}(A) \quad \text{for all} \quad j = 1, \dots, n."$$

Of course, also this version may be applied to all the subjects mentioned in Section 2.

D) The Approximation Lemma remains valid with the same number k of positive q_i if the matrix A is blown up by introducing “new” columns that are merely convex combinations of already existing columns.

E) The Approximation Lemma is not an isolated discovery. In several fields results of a similar flavour have been obtained, for instance:

(i) the VC dimension [VC 71] and its applications in computational learning theory [LLW 91], [AB 92] and computational geometry [HW 87],

(ii) Monte Carlo approximations [KL 83], [KLM 89], [LV 91],

(iii) uniformity and irregularities (discrepancies) of set systems and matrices [BF 81], [LSV 86],

(iv) stochastic information theory [Ahl 78], [Ahl 79].

On the other hand there are some statistics papers by Wald and others [DWW 51], [WW 51] with similarly sounding titles but different contents.

We conclude with an open problem.

F) A pair (p, A, ε) consisting of a probability vector $p = (p_1 \dots p_m)$, a matrix $A \in [0,1]^{m \times n}$, and a constant $\varepsilon > 0$ is called (m, n, s)*–extremal* for some $s \in \mathbb{N}$ if two statements hold:

(i) There is no ε–approximation q to (p, A) with at most s non–negative entries q_i .

(ii) For every probability vector $p' = (p'_1 \dots p'_m)$ and every matrix $A' \in [0,1]^{m \times n}$ there is an ε–approximation q' to (p', A') with at most $(s+1)$ non–negative entries q'_i .

Question: Find or characterize (n, m, s)–extremal tuples (p, A, ε) for small parameters n, m , and s .

Acknowledgement

My work on the Approximation Lemma was influenced and stimulated by discussions with several people.

Especially, I would like to thank Rudolf Ahlswede, Mike Luby, Seffi Naor, Sören Perrey, Professor Joachim Rosenmüller and his group, Torsten Sillke, Uli Tamm, Hans Zessin, and the students in the “data compression” course which I gave at the University of Bielefeld in the winter term 92/93.

Thanks are also due to an anonymous referee for his constructive criticism.

References

[AB 92] Anthony, M. and Biggs, N. 1992. Computational Learning Theory: An Introduction. Cambridge, UK: Cambridge University Press.

[Ahl 78] Ahlswede, R. 1978. Elimination of correlation in random codes for arbitrarily varying channels. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 44: 159–175.

[Ahl 79] Ahlswede, R. 1979. Colouring hypergraphs: a new approach to multi–user source coding, part I. Journal of Combinatorics, Information Theory, and System Sciences 4: 76–115.

[Alt 93a] Althöfer, I. 1993. On sparse approximations to randomized strategies and convex combinations, SFB 343 "Diskrete Strukturen in der Mathematik", Universität Bielefeld, Preprint 93–004.

[Alt 93b] Althöfer, I. 1993. On shortening of linear evaluation functions. Submitted to International Computer Chess Association Journal.

[ASE 92] Alon, N., Spencer, J., and Erdös, P. 1992. The Probabilistic Method. New York, NY: Wiley.

[BF 81] Beck, J. and Fiala, T. 1981. Integer–making theorems. Discrete Applied Mathematics 3: 1–8.

[CT 80] Condon, J.H. and Thompson, K. 1984. Belle, in "Chess Skill in Man and Machine" (P.W. Frey, editor). New York, N.Y.: Springer.

[CT 91] Cover, T.M. and Thomas, J.A. 1991. Elements of Information Theory. New York, N.Y.: Academic Press.

[DWW 51] Dvoretzky, A., Wald, A., and Wolfowitz, J. 1951. Elimination of randomization in certain statistical decision procedures and zero–sum two–person games. Annals of Mathematical Statistics 22: 1–21.

[ES 74] Erdös, P. and Spencer, J. 1974. Probabilistic Methods in Combinatorics. New York, N.Y.: Academic Press.

[Hoe 63] Hoeffding, W. 1963. Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association 58: 13–30.

[HW 87] Haussler, D. and Welzl, E. 1987. Epsilon–nets and simplex range queries. Discrete and Computational Geometry 2: 127–151.

[Kar 59] Karlin, S. 1959. Matrix Games, Programming, and Mathematical Economics. Reading, MA: Addison–Wesley.

[KL 83] Karp, R.M. and Luby, M. 1983. Monte–Carlo algorithms for enumeration and reliability problems, in Proc. 24th IEEE Symp. FoCS: 56–64.

[KLM 89] Karp, R.M., Luby, M., and Madras, N. 1989. Monte Carlo approximation algorithms for enumeration problems. Journal of Algorithms 10: 429–448.

[KY 76] Knuth, D.F. and Yao, A.C. 1976. The complexity of random number generation, in Algorithms and Complexity: Recent Results and New Directions (Traub, J.F., Editor), New York, N.Y.: Academic Press.

[LLW 91] Littlestone, N., Long, P.M., and Warmuth, M.K. 1991. On–line learning of linear functions, in Proc. 23rd ACM Symposium on Theory of Computing, New Orleans: 465–475.

[LSV 86] Lovasz, L., Spencer, J.H., and Vesztergombi, K. 1986. Discrepancy of set systems and matrices. European Journal of Combinatorics 7: 151–160.

[LV 91] Luby, M. and Velickovic, B. 1991. On deterministic approximation of DNF, in Proc. 23rd ACM Symposium on Theory of Computing, New Orleans: 430–438.

[McD 89] McDiarmid, C. 1989. On the method of bounded differences, pp. 148–188 in Surveys in Combinatorics, London Mathematical Society Lecture Note Series 141 (J. Siemons, Ed.), Cambridge, UK: Cambridge University Press.

[Ros 81] Rosenmüller, J. 1981. The Theory of Games and Markets. Amsterdam, NL: North Holland.

[Sha 50] Shannon, C.E. 1950. Programming a computer for playing chess. Philosophical Magazine 41: 256–275.

[VC 71] Vapnik, V.N. and Chervonenkis, A.Ya. 1971. On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications 16: 264–280.

[WW 51] Wald, A. and Wolfowitz, J. 1951. Two methods of randomization in statistics and the theory of games. Annals of Mathematics 53: 581–586.