EDGE SEARCH IN HYPERGRAPHS

Ingo Althöfer and Matthias Löwe

Fakultät für Mathematik
Universität Bielefeld
Postfach 100131
33501 Bielefeld
Germany

Abstract

Consider an undirected hypergraph $H = (X, E)$ with a probability distribution P on the set E of its hyperedges. We investigate the average case complexity $\overline{L}(H, P)$ of finding an unknown hyperedge $e^* \in E$, chosen according to P, if only tests are allowed that check, whether e^* is contained in the induced subhypergraphs $H[Y]$ for $Y \subset X$, or not.

In this note we prove that for hypergraphs with average size of the hyperedges r $H(P) \leq \overline{L}(H, P) < H(P) + 3r$. Here $H(\cdot)$ denotes the usual entropy function. The lower bound is the information theoretic bound, whereas the upper bound stems from a construction presented in this note.

More precisely, our construction gives search lengths which are bounded from above by $-\log_2 P(e^*) + 3 \cdot |e^*|$ for all possible choices of e^*.

1. Introduction

In [Aig 86] the following search problem on graphs was introduced which generalizes the group testing problem with two defectives.
Given a finite graph $G = (X, E)$ without loops and multiple edges, we search for an edge $e^* \in E$. Admissible tests are the following: For any subset $Y \subset X$ we may test whether e^* is contained in the induced subgraph $G[Y]$ or not. For the worst case complexity $L(G)$ of this problem the information theoretic lower bound is $L(G) \geq \lceil \log_2 |E| \rceil$ and there are graphs with $L(G) = \lceil \log_2 |E| \rceil + 1$. Recently Damaschke [Dam 93] proved that indeed for all graphs G

$$L(G) \leq \lceil \log_2 |E| \rceil + 1.$$

In this note we generalize the above problem in two ways. First we consider hypergraphs $H = (X, E)$ instead of graphs. Second we investigate the average search length in case of a probability distribution $P = \big(P(e_1) \ldots, P(e_m)\big)$ $(m = |E|)$ on the set E of hyperedges. For a fixed search scheme and every $e \in E$ we denote by cost(e) the number of questions required in case of $e^* = e$. For given H and P let $\overline{L}(H, P)$ be the minimum possible value of the average search cost $\sum_{e \in E} P(e)\text{cost}(e)$. Recall, that the information theoretic lower bound is

$$\overline{L}(H, P) \geq H(P)$$

where $H(P) = -\sum_{i=1}^{m} P(i) \log_2 P(i)$ denotes the entropy of P. Naively we might hope to derive an upper bound for $\overline{L}(H, P)$ such that $\overline{L}(H, P) \leq H(P) + c$ with a constant c independent of H (especially with c independent of $|E|$). The example, where $|X| = n$, E consists of all hyperedges of size $n - 1$, and P is the uniform distribution on E, such that the average search cost are obviously $\frac{n}{2}$, shows that this hope is not very sensible without any further requirements.
In fact we are able to give such an upper bound on $\overline{L}(H, P)$ under the weak hypothesis that the average size of the hyperedges is known, i.e.

$$\sum_{e \in E} P(e)|e| = r \tag{1}$$

for some r (here $|e|$ stands for the number of vertices in e). Indeed we prove

Theorem 1. *For every hypergraph $H = (X, E)$ with a probability distribution P on E there is a search scheme yielding*

$$\textit{cost}(e) \leq -\log_2 P(e) + 3|e| \quad \textit{for every} \quad e \in E.$$

An immediate consequence is:

Corollary 2. *For every hypergraph H and every probability distribution P on E such that (1) holds we get*

$$\overline{L}(H, P) \leq H(P) + 3r.$$

In the case of equal distribution on the set of hyperedges of an hypergraph with bounded rank r Theorem 1 yields a new and simpler proof the Hypergraph Theorem 3.4 in [AlT 93, p. 8]. Moreover the constant c_r is now specified as $3r$. As will be explained at the end of the next section the constant $3r$ can be improved for r–uniform hypergraphs, i.e. for hypergraphs with uniform size of the hyperedges $|e| = r$.

Corollary 3. *For every r–uniform hypergraph H and every probability distribution P on E there is a search scheme with*

$$cost(e) \leq -\log_2 P(e) + 2r \quad \textit{for every} \quad e \in E$$

such that

$$\overline{L}(H, P) \leq H(P) + 2r.$$

Especially if H is an ordinary, undirected graph without loops and multiple edges we have

$$\overline{L}(H, P) \leq H(P) + 4.$$

This generalizes the result $\lceil \log_2 |E| \rceil + 1$ of [Dam 93] to weighted graphs.
In the following we present the search scheme on which Theorem 1 is based and prove Theorem 1.
We refer the reader to [Aig 88] for an introduction to combinatorial search theory and to [DH 93] for a rich survey on group testing problems. In [AW 87] search problems are discussed in a more general setting.

2. Presentation of the Algorithm and Proof of Theorem 1

Our algorithm is based on the so-called alphabetic search. Let $Z = (z_1, \dots, z_n)$ be an ordered set with $z_i < z_j$ whenever $i < j$ and π be a probability distribution on Z. A search problem on (Z, π) is called an alphabetic search problem, if only tests of the form "$z^* \geq z_j$?", $1 \leq j \leq r$ are allowed. It is well known (see [AW 87, Thm. 4.3.5]) that for such a setting there exists a search scheme such that the search length of finding an unknown z^* is less than $-\log_2 \pi(z^*) + 2$ for all z^*. For a more detailed discussion of alphabetic search see [AW 87, Chapter 4].
Let $H = (X, E)$ be the given hypergraph and $e^* \in E$ be the unknown edge chosen according to a probability distribution P on E. The following algorithm consists of several steps, each step itself consisting of one or two parts. In the first part of each step we determine one vertex in e^*, whereas in the second part we check, whether we have already found e^* completely.
To be more precise: Let us take any enumeration of $X = (x_1, \dots, x_n)$ and the natural order belonging to this enumeration (i.e. $x_i < x_j$ whereas $i < j$). For convenience let us agree upon notating any hyperedge $e = (e_1, \dots, e_s) \in E$ in ascending order, i.e. $e_1 < e_2 < \dots < e_s$. Moreover let $e^* = (e_1^*, \dots, e_r^*)$. For $t \leq r$ we denote by E_t the set of all hyperedges remaining before the t'th step of the search. Especially, we have $E_1 = E$. We will see that E_t consists of all $e \in E$ with $|e| \geq t$ and $e_j = e_j^*$ for $1 \leq j \leq t-1$. Now we define a search scheme on E_t in the following way.
Part I: Let P_t be the probability distribution P conditioned on E_t, i.e. $P_t(e) = \frac{P(e)}{P(E_t)}$ if $e \in E_t$ and $P_t(e) = 0$ otherwise. By $Q_t(x)$ we denote the probability measure on X induced by P_t via the relation that x is the next vertex of an hyperedge in E_t. More

formally: $Q_t(x) = \sum_{\substack{e \in E_t \\ e_t = x}} P_t(e)$. Now we search alphabetically on (X, Q_t) for e_t^* where we use a search scheme such that for the search length $\ell(\cdot)$ it holds

$$\ell(e_t) \leq \log_2 Q_t(e_t) + 2$$

in all cases (see above).
Of course, we have to check, whether this alphabetic search is compatible with the admissible questions. We can answer this question in the affirmative, since the test $x \geq x_j$ (with $e_{t-1}^* \leq x_j \leq x_s$) in the alphabetic search scheme corresponds to the admissible question $e^* \subset H[Y]$ where $Y = \{e_1^*, \ldots, e_{t-1}^*\} \cup \{x \in X : x \geq x_j\}$ (and $H[Y]$ is the induced hypergraph on Y) in the hypergraph search.
Part II: This part of a step only becomes active if the hyperedge $\{e_1^*, \ldots, e_t^*\}$ belongs to E. Then we check, if we are already done, i.e. we perform the test $e^* = \{e_1^*, \ldots, e_t^*\}$. Obviously the search ends after $r = |e^*|$ steps. Hence we are able to prove Theorem 1:
Proof of Theorem 1: As mentioned in the explanation of the algorithm the search length of the first part of the t'th step is less than $-\log Q_t(e_t^*) + 2$. Hence the total search costs are bounded by

$$\mathrm{cost}(e^*) \leq -\log \prod_{t=1}^{r} Q_t(e_t^*) + 3r.$$

Since

$$-\log \prod_{t=1}^{r} Q_t(e_t^*) = -\log \left[\frac{\sum_{e: e_1 = e_1^*} P(e)}{1} \cdot \frac{\sum_{e: e_1 = e_1^*, e_2 = e_2^*} P(e)}{\sum_{e: e_1 = e_1^*, |e| > 1} P(e)} \cdots \frac{\sum_{e: e_1 = e_1^*, \ldots, e_r = e_r^*} P(e)}{\sum_{e: e_1 = e_1^*, \ldots, e_{r-1} = e_{r-1}^*} P(e)} \right]$$

$$\leq -\log \left[\sum_{\substack{e: \\ e_1 = e_1^*, \ldots, e_r = e_r^*}} P(e) \right] \leq -\log P(e^*)$$

we have

$$\mathrm{cost}(e^*) \leq -\log P(e) + 3r$$

proving our Theorem. (The argument above is similar to the one on which the grouping property of the entropy function is based.)

□

Note that the constants $c_r = 2r$ and specially $c_2 = 4$ in Corollary 3 instead of $3r$ (resp. 6) stem from the fact that in the case of r–uniform hypergraphs there is no need to perform Part II of a search step at any time.

Let us conclude with a remark asking for two possible improvements of our result.

Remarks 4.

(i) *Let us step back to the example mentioned in the introduction. Let* $|X| = n$ *and* E *be the set of all hyperedges* $e \subset X$ *with* $|e| = n-1$*. Let* P *be the uniform distribution on* E*. This example indicates that the best possible constant* $c(e)$ *such that* $cost(e) \leq -\log P(e) + c(e)$ *should not be smaller than* $\frac{|e|}{2}$*. It remains open, whether this lower bound can actually be achieved.*

(ii) *Theorem 1 gives a good bound for every hypergraph, in the sense that the natural candidate* $-\log P(e)$ *from the information theoretic lower bound is missed by at most* $3|e|$*. Such a statement for every edge does not hold for several traditional codes, including the optimal Huffmann code: Assume the distribution* $(p_1, \ldots, p_m)$ *with* $\frac{p_{i+1}}{p_i} = \frac{\sqrt{s}-1}{2}$ *for all* i*. In the Huffmann code the i'th codeword gets length* $\ell_i = i$*, but* $p_i \approx \left(\frac{\sqrt{5}-1}{2}\right)^i$*. Hence*

$$\ell_i \approx \frac{1}{\log_2 \frac{2}{\sqrt{5}-1}} \cdot \log_2 \frac{1}{p_i} \approx 1.44 \log_2 \frac{1}{p_i}.$$

Therefore one may be interested in, whether we can improve the constant in Corollary 2 by considering such codes that violate $cost(e) \leq -\log P(e) + 3 \cdot |e|$ *for some* $e \in E$*. Moreover we may ask for the best possible constant for the average case complexity and compare it with the best possible constant in Theorem 1.*

References

[Aig 86] M. Aigner. Search problems on graphs. Discrete Applied Mathematics 14 (1986), 215–230.

[Aig 88] M. Aigner. Combinatorial Search. Wiley–Teubner, New York, 1988.

[AiT 93] M. Aigner and E. Triesch. Realizability and Uniqueness in Graphs. To appear in Discrete Mathematics.

[AlT 93] I. Althöfer and E. Triesch. Edge search in graphs and hypergraphs of bounded rank. Discrete Mathematics 115 (1993), 1–9.

[AW 87] R. Ahlswede and I. Wegener. Search Problems. Wiley–Teubner, New York, 1987. (Translation of the original German version “Suchprobleme”, Teubner, 1979.)

[Bol 79] B. Bollobas. Graph Theory: An Introductory Course. Springer, New York, 1979.

[Dam 93] P. Damaschke. A tight upper bound for group testing in graphs. To appear in Discrete Mathematics.

[DH 93] D. Du and F. Hwang. Combinatorial Group testing and its Applications. World Scientific Publications, Singapure (1993).

[Tri 90] E. Triesch. Suchprobleme in Graphen und Hypergraphen. Habilitation Thesis, RWTH Aachen, Department of Mathematics and Computer Science.