SUBSET CAPACITY OF FINITE SETS AND APPLICATIONS IN COMPUTATIONAL NEUROSCIENCE

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Computer Science

by

Chandradeep Chowdhury

December 2023

COMMITTEE MEMBERSHIP

TITLE:	Subset Capacity of Finite Sets and Applications in Computational Neuroscience
AUTHOR:	Chandradeep Chowdhury
DATE SUBMITTED:	December 2023
COMMITTEE CHAIR:	Mugizi Robert Rwebangira, Ph.D. Professor of Computer Science
COMMITTEE MEMBER:	Rodrigo De Moura Canaan, Ph.D. Professor of Computer Science
COMMITTEE MEMBER:	Theresa Anne Migler, Ph.D. Professor of Computer Science

ABSTRACT

Subset Capacity of Finite Sets and Applications in Computational Neuroscience

Chandradeep Chowdhury

Your abstract goes in here

ACKNOWLEDGMENTS

Thanks to:

- My parents, and grandparents.
- My thesis committee members.
- My collaborators, Patrick Perrine, and Shosei Anegawa.
- Cal Poly Graduate Education Office, for supporting me with a tuition waiver for academic year 2022-2023
- Cal Poly Cares, for supporting me with a Grant.
- Cal Poly Housing Administration, for supporting me with emergency housing for two quarters.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

Chapter 2

RESULTS

2.1 Interference

We now formally define a notion of interference.

Definition 1. (***k*-Interference**) Given two sets U, W, and some number $k \in (0, |W|]$, we say U $k-$*interferes* with W if

$$|U \cap W| \geq \frac{|W|}{k}. \tag{2.1}$$

This is a generalization of the notion of interference introduced by Valiant in 2005. Valiant defines a memory to be in a "firing" state if more than half the nodes in the memory are in a "firing" state. He then defines interference as the unintential firing of a memory W when another memory U is fired, which is possible if and only if more than half the nodes of W are also present in U [1]. This corresponds to the $k = 2$ case of our definition.

We are now interested in finding the probability of a randomly picked subset interfering with another randomly picked subset. We start with the case where they are randomly picked as we believe it is the simplest case.

Lemma 2. *Given a set V with n items and two subsets U, W of respective sizes r_u, r_w, denote the size of the intersection between them by the random variable Y. Then the* *probability of U k-interfering with W is*

$$\sum_{y=\lceil \frac{r_w}{k} \rceil}^{r_w} \frac{\binom{r_u}{y}\binom{n-r_u}{r_w-y}}{\binom{n}{r_w}}$$

and $Y \sim Hypergeometric(n, r_u, r_w)$.

Proof. If $V = \{v_1, ..., v_n\}$, we can represent the first randomly picked subset U as a boolean vector u of length n defined by

$$u_i = \begin{cases} 1 & \text{if } v_i \in U \\ 0 & \text{if } v_i \notin U. \end{cases}$$

With this representation, U will intersect another randomly picked subset W at the indices where both boolean vectors u, w have a 1. Then Y denotes the number of indices where both u, w have a 1. First note that

$$\mathbb{P}(Y = y) = \frac{\binom{r_u}{y}\binom{n-r_u}{r_w-y}}{\binom{n}{r_w}}. \tag{2.2}$$

This follows from the fact that given the first vector U, we already know where the 1's are located. We can pick the y intersecting 1's for the second vector in $\binom{r_u}{y}$ ways implicitly placing 0's in the remaining spots. We then fill the remaining $n - r_u$ indices corresponding to the 0's in the first vector with $r_w - y$ 1's in $\binom{n-r_u}{r_w-y}$ ways. Finally we divide by the total number of possible subsets $\binom{n}{r_w}$. Clearly, this is the probability mass function of the hypergeometric distribution with population size n, r_u success states and r_w draws. We conclude that $Y \sim Hypergeometric(n, r_u, r_w)$. Finally, to find the probability of U k-interfering with W we need to find $\mathbb{P}(Y \geq \lceil \frac{r_w}{k} \rceil)$ which is the sum of $\mathbb{P}(Y = y)$ from $y = \lceil \frac{r_w}{k} \rceil$ to $y = r_w$. □

For brevity, we can reinterpret the above probability as the tail distribution function of Y at $\left\lfloor \frac{r_w}{k} \right\rfloor$,

$$\mathbb{P}\left(Y \geq \left\lceil \frac{r_w}{k} \right\rceil\right) = \mathbb{P}\left(Y > \left\lfloor \frac{r_w}{k} \right\rfloor\right) = \bar{F}_Y\left(\left\lfloor \frac{r_w}{k} \right\rfloor\right)$$

Recall from statistics that the expectation of a binary payoff, like intersection, that depends on a cutoff (in this case $\left\lceil \frac{r_w}{k} \right\rceil$) is equal to $\mathbb{P}\left(Y \geq \left\lceil \frac{r_w}{k} \right\rceil\right)$. Therefore the probability in lemma 2 is equal to the expected number of interferences of U with W.

We then want to find an expectation for the expected number of interferences when the sizes of the subsets are derived from a distribution with a common mean. This approach will make our results more applicable to models like the Neuroidal Model that assume memory sizes follow some distribution [1].

Finding this expectation without any further assumptions is quite hard as the expectation operator does not behave nicely with binomial coefficients as in lemma 2. Instead we will make a few reasonable assumptions that will allow us to derive a reasonable lower bound for this expectation, especially as the size of set $n \to \infty$. Before proceeding further, we will establish a well-known lower and upper bound for the binomial coefficent.

Lemma 3. *Given $n \in [0, \infty)$ and $k \in (0, \infty)$, we have*

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k$$

Proof. First we will prove the lower bound. For $k = 1$, we have

$$\left(\frac{n}{k}\right)^k = n \leq n = \binom{n}{k}.$$

Then, for $k > 1$ pick m such that $0 < m < k \leq n$. Then we have

$$k \leq n \Rightarrow \frac{m}{n} \leq \frac{m}{k} \Rightarrow 1 - \frac{m}{k} \leq 1 - \frac{m}{n} \Rightarrow \frac{k-m}{k} \leq \frac{n-m}{n} \Rightarrow \frac{n}{k} \leq \frac{n-m}{k-m}$$

Therefore,

$$\left(\frac{n}{k}\right)^k = \frac{n}{k} \cdot \ldots \cdot \frac{n}{k} \leq \frac{n}{k} \cdot \frac{n-1}{k-1} \cdot \ldots \cdot \frac{n-k+1}{1} = \binom{n}{k}.$$

Now let us consider the upper bound. Recall the expansion of e^k,

$$e^k = \sum_{j=0}^{\infty} \frac{k^j}{j!}$$

. Simply picking the $k+1$-th term only we get,

$$e^k \geq \frac{k^k}{k!}$$

which implies

$$\frac{1}{k!} \leq \left(\frac{e}{k}\right)^k$$

Then observe that,

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} = \frac{n \cdot (n-1) \ldots (n-(k-1))}{k!} \leq \frac{n^k}{k!} \leq \left(\frac{en}{k}\right)^k$$

□

We will now use these bounds to find the desired expectation.

Lemma 4. *Given a set V with n items and two subsets U, W of respective sizes r_u, r_w, denote the size of the intersection between them by the random variable Y. If*

1. *r_u, r_w are random variables that follow distributions with expected value r,*

2. $r_u, r_w \in [r-\delta, r+\delta]$ *for some* $\delta > 0$*,*

3. $n > 2r + \delta$

4. $Y \sim Hypergeometric(n, r_u, r_w)$*,*

then

$$\mathbb{E}\left[\bar{F}_Y\left(\left\lfloor \frac{r_w}{k} \right\rfloor\right)\right] \geq \sum_{y=\left\lceil \frac{r-\delta}{k} \right\rceil}^{\lfloor r-\delta \rfloor} \frac{\left(\frac{r}{y}\right)^y \left(\frac{n-r}{r+\delta-y}\right)^{r-\delta-y}}{\left(\frac{en}{r}\right)^{r+\delta}}$$

Remark. Before proceeding with the proof, we want to justify the assumptions made here. The first and fourth assumption are necesary as that is the context in which we think this result will be the most useful. The second assumption is simply bounding the subset size by a particular offset from r that can be varied and smaller the value of δ, the better the tighter the lower bound of the expectation will be. This is necessary to obtain a clean result in terms of r only without involving the rnadom variables r_w, r_u that are specific to particular subsets. The third assumption is simply saying that n should be significantly bigger than r which is often the case in our intended field of application [1]. We chose to go with a tight inequality instead of a more general statement like $n >> r$ in order to make the result as strong as possible. This is also necessary for a critical step in the proof.

Proof. First note that $n > r_u, r_w$ and by extension $n > r$ since the size of a subset cannot exceed the size of the set. Then observe that

$$\begin{aligned}
\mathbb{E}\left[\bar{F}_Y\left(\left\lfloor \frac{r_w}{k} \right\rfloor\right)\right] &= \mathbb{E}\left[\sum_{y=\left\lceil \frac{r_w}{k} \right\rceil}^{r_w} \mathbb{P}(Y = y)\right] \\
&= \mathbb{E}\left[\sum_{y=\left\lceil \frac{r_w}{k} \right\rceil}^{r_w} \frac{\binom{r_u}{y}\binom{n-r_u}{r_w-y}}{\binom{n}{r_w}}\right] \\
&= \sum_{y=\left\lceil \frac{r_w}{k} \right\rceil}^{r_w} \mathbb{E}\left[\frac{\binom{r_u}{y}\binom{n-r_u}{r_w-y}}{\binom{n}{r_w}}\right] \\
&\geq \sum_{y=\left\lceil \frac{r_w}{k} \right\rceil}^{r_w} \mathbb{E}\left[\frac{\left(\frac{r_u}{y}\right)^y \left(\frac{n-r_u}{r_w-y}\right)^{r_w-y}}{\left(\frac{en}{r_w}\right)^{r_w}}\right] \\
&= \sum_{y=\left\lceil \frac{r_w}{k} \right\rceil}^{r_w} \frac{\left(\frac{r}{y}\right)^y \left(\frac{n-r}{r-y}\right)^{r_w-y}}{\left(\frac{en}{r}\right)^{r_w}} \\
&\geq \sum_{y=\left\lceil \frac{r_w}{k} \right\rceil}^{r_w} \frac{\left(\frac{r}{y}\right)^y \left(\frac{n-r}{r+\delta-y}\right)^{r_w-y}}{\left(\frac{en}{r}\right)^{r_w}} \\
&\geq \sum_{y=\left\lceil \frac{r-\delta}{k} \right\rceil}^{\lfloor r-\delta \rfloor} \frac{\left(\frac{r}{y}\right)^y \left(\frac{n-r}{r+\delta-y}\right)^{r_w-y}}{\left(\frac{en}{r}\right)^{r_w}} \\
&\geq \sum_{y=\left\lceil \frac{r-\delta}{k} \right\rceil}^{\lfloor r-\delta \rfloor} \frac{\left(\frac{r}{y}\right)^y \left(\frac{n-r}{r+\delta-y}\right)^{r-\delta-y}}{\left(\frac{en}{r}\right)^{r+\delta}}
\end{aligned} \tag{2.3}$$

The first and second equalities follow from the definition of the tail distribution and PMF of the hypergeometric distribution respectively. The third equality follows from the linearity of the expectation operator. The fourth inequality follows from applying the lower bound and upper bound we established in lemma 3 to the numerator and demonimation respectively as well as the fact that the expectation operator preservers order. The fifth equality follows from the expectation operator preserving multiplication. The sixth inequality follows since increasing the denominator slightly results in a smaller value. The seventh inequality follows from the second assumption and since all the terms in the expression are positive and reducing the number of terms we sum over will reduce the total value. Note that even though going from $\lceil \frac{r_w}{k} \rceil$ to $\lceil \frac{r-\delta}{k} \rceil$ might add a few more terms, it will always be less than the terms removed by switching the upper limit from r_w to $\lfloor r - \delta \rfloor$. The eighth inequality follows from the second and third assumptions, since $n > 2r + \delta \implies n > 2r + \delta - y \implies n - r > r + \delta - y$ and $n > 2r + \delta \implies en > r$ so both the terms are positive and we can replace r_w with its lower and upper bound respectively. □

2.2 Capacity

With the above lemmas in our arsenal we can now move on the main subject of this thesis.

BIBLIOGRAPHY

[1] L. G. Valiant. Memorization and Association on a Realistic Neural Model. Neural Computation, 17(3):527–555, 2005.

[2] L. G. Valiant. Capacity of Neural Networks for Lifelong Learning of Composable Tasks. In 58th Annual Symposium on Foundations of Computer Science, pages 367–378. IEEE, 2017.