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

Page

CHAPTER

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

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

Corollary 2. *If $|U| = |W|$, then U k-interferes with W if and only if W k-interferes with U.*

We restrict the upper range of k to $|W|$ for convenience, as beyond that all values of $\frac{|W|}{k}$ will be less than 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 3. *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 $y = r_w$. $\square$

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

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

Recall from statistics that the expectation of a binary payoff, like intersection, that depends on a cutoff (in this case $\lceil \frac{r_w}{k} \rceil$) is equal to $\mathbb{P}\left(Y \geq \lceil \frac{r_w}{k} \rceil\right)$. Therefore the probability in lemma 3 is equal to the expected number of interferences of U with W.

We then want to estimate the expected number of interferences when the sizes of the subsets are within a certain offset of r, say δ without being exactly equal to r. This approach will make our results more applicable to models like the Neuroidal Model that assume memory sizes follow some distribution [1]. The offset can be selected to best suit the distribution involved. For example if the sizes come from $\mathcal{N}(r, \sigma)$, it makes sense to choose $\delta = 3\sigma$ since 99.7% of all values lie within $[r - 3\sigma, r + 3\sigma]$.

Generalizing this without any further assumptions is quite hard as the expectation operator does not behave nicely with binomial coefficients as in lemma 3. Instead we will make a reasonable assumption that will allow us to derive a reasonable lower bound for this expectation in terms of a general parameter instead of individual subset sizes.

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

2. $n >> 2(r + \delta)$,

then

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

Remark. Before proceeding with the proof, we want to justify the second assumption made here. It is a known fact that bounding binomial coefficients above or below is hard due to the nature of how it varies with respect to the second argument. We know that $\binom{n}{k}$ reaches its maximum value at $\left\lceil \frac{n}{2} \right\rceil$ or $\left\lfloor \frac{n}{2} \right\rfloor$ and it is monotonically increasing at smaller values and decreasing at larger values. My making the assumption here we can ensure that our second argument is always a lot smaller than this maxima, and as such an increase in the second argument will only increase the value of the expression. This assumption is reasonable since models like the Neuroidal Model expect the memory sizes to be significantly smaller than the size of the model [1]. Also note that the binomial coefficient increases monotonically with respect to the first argument.

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

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

The first and second equalities follow from the definition of the tail distribution and lemma 3 respectively. The second inequality follows from assumption 1. in the theorem and the behavior of the binomial coefficient under varying arguments. The final inequality follows from the fact that since all terms in the sum are positive, reducing the number of terms will make the overall expression smaller. Note that while reducing the lower limit from $\lceil \frac{r_w}{k} \rceil$ to $\lceil \frac{r-\delta}{k} \rceil$ might lead to addition of more terms, it will always be less than the number of terms removed due to changing the upper limit from r_w to $r - \delta$.

□

2.2 Capacity

With the above lemmas in our arsenal we can now move on the main subject of this thesis. We now formally define the capacity of a system of overlapping subsets with interference being the limiting factor.

Definition 5. (**(r, T, k, δ)-Subset Capacity**) Given a set $V = \{v_1, ..., v_n\}$, and parameters $r, T, k, \delta > 0$, the (r, T, k, δ)-*subset capacity* of V is the *maximum* number of subsets that can be picked subject to the conditions that any randomly picked subset U,

1. $r' \in [r - \delta, r + \delta]$,
2. $n >> 2(r + \delta)$,
3. $E[X_U] \leq T$ where X_U is a random variable denoting the number of interferences caused due to picking U.

We need the second restriction on the memories here since we want to apply lemma 4 to every pair. The third restriction here can be thought of as a stopping criteria as we stop picking the subsets once the expectation of interference reaches that threshold. In the context of models in computational neuroscience like the Neuroidal Model, this means that there will be too impact on the quality of memorization, that is too much noise and misfiring in the system if we add any further memories.

Before deriving the capacity for the general case, let us consider the simpler case where all memories have the exact same size. This is valuable since it results in a much simpler expression and we can use this as an approximation for the more general case too. However note that we realize that this scenario is not biologically plausible at all.

Theorem 6. *Given a set V with n items and the property that every picked subset will have size exactly r, the (r, T, k, δ)-subset capacity of V is*

$$\left\lfloor \frac{T}{\bar{F}_Y\left(\left\lfloor \frac{r}{k} \right\rfloor\right)} + 1 \right\rfloor .$$

Remark. Since all subsets have fixed size r, note that the choice of δ is not relevant here.

Proof 1. Suppose we have already have $M - 1$ subsets in the universe. Pick a random subset U. From lemma 3, we know that the expected number of k-interferences of U with another arbitrary subset W from the universe is $\bar{F}_Y\left(\left\lfloor \frac{r}{k} \right\rfloor\right)$. Since there are $M - 1$ other subsets, the total expected number of k-interferences caused by picking U is $(M - 1)\bar{F}_Y\left(\left\lfloor \frac{r}{k} \right\rfloor\right)$.

From inequality 3 in the definition of capacity, we have

$$(M - 1)\bar{F}_Y\left(\left\lfloor \frac{r}{k} \right\rfloor\right) \leq T \implies M \leq \frac{T}{\bar{F}_Y\left(\left\lfloor \frac{r}{k} \right\rfloor\right)} + 1. \tag{2.4}$$

The (r, T, k, δ)-subset capacity of V then is the largest integer M that satisfies inequality 2.4. □

We provide an alternate proof that, while less elegant, can be scaled to prove the general statement.

Proof 2. Suppose we have already have M subsets in the universe. Pick two subsets U, W without replacement. From lemma 3, we know that the expected number of k-interferences of U with W is $\bar{F}_Y\left(\left\lfloor \frac{r}{k} \right\rfloor\right)$. Since we know all subsets have the same size, the expected number of k-interferences of W with U is the same. So the expected number of interferences caused by one pair is

$$2\bar{F}_Y\left(\left\lfloor\frac{r}{k}\right\rfloor\right).$$

We know that there are $\binom{M}{2} = M(M-1)/2$ such pairs so the expected number of total interferences is

$$2 \cdot \frac{M(M-1)}{2}\bar{F}_Y\left(\left\lfloor\frac{r}{k}\right\rfloor\right) = M(M-1)\bar{F}_Y\left(\left\lfloor\frac{r}{k}\right\rfloor\right).$$

Since there are M subsets, the expected number of interferences by choosing picking one subset is

$$\frac{M(M-1)}{M}\bar{F}_Y\left(\left\lfloor\frac{r}{k}\right\rfloor\right)) = (M-1)\bar{F}_Y\left(\left\lfloor\frac{r}{k}\right\rfloor\right).$$

From inequality 3, we have

$$(M-1)\bar{F}_Y\left(\left\lfloor\frac{r}{k}\right\rfloor\right) \leq T \implies M \leq \frac{T}{\bar{F}_Y\left(\left\lfloor\frac{r}{k}\right\rfloor\right)} + 1. \tag{2.5}$$

The (r, T, k, δ)-subset capacity of V is the largest integer M that satisfies inequality 2.5. □

We will now tackle the general case using the same strategy as above.

Theorem 7. *Given a set V with n items, the (r, T, k, δ)-subset capacity of V is bounded above by*

$$\frac{T}{\sum_{y=\lceil\frac{r-\delta}{k}\rceil}^{\lfloor r-\delta\rfloor}\frac{\binom{r-\delta}{y}\binom{n-r-\delta}{r-\delta-y}}{\binom{n}{r+\delta}}} + 1$$

Remark. Note that we can only say it is bounded above and not the exact capacity as defined since we have to use lemma 4. However as $\delta \to 0$, this expression converges to the expression in theorem 6.

Proof. Suppose we have M subsets $U_1, ..., U_M$ with sizes $r_1, ..., r_M$. Pick two subsets U_i, U_j. From lemma 3, we know that the expected number of interferences caused by this pair is

$$\bar{F}_Y\left(\left\lfloor \frac{r_j}{k} \right\rfloor\right) + \bar{F}_Y\left(\left\lfloor \frac{r_i}{k} \right\rfloor\right).$$

We then sum over all possible pairings to get the expected number of total interferences:

$$\sum_{(i,j)\in\mathbb{Z}\times\mathbb{Z},1\leq i,j\leq M,i\neq j} \left(\bar{F}_Y\left(\left\lfloor \frac{r_j}{k} \right\rfloor\right) + \bar{F}_Y\left(\left\lfloor \frac{r_i}{k} \right\rfloor\right)\right).$$

Since there are M subsets, the expected number of interferences by picking one subset is

$$\frac{1}{M}\sum_{(i,j)\in\mathbb{Z}\times\mathbb{Z},1\leq i,j\leq M,i\neq j} \left(\bar{F}_Y\left(\left\lfloor \frac{r_j}{k} \right\rfloor\right) + \bar{F}_Y\left(\left\lfloor \frac{r_i}{k} \right\rfloor\right)\right).$$

From inequality 3, we have

$$\frac{1}{M}\sum_{(i,j)\in\mathbb{Z}\times\mathbb{Z},1\leq i,j\leq M,i\neq j} \left(\bar{F}_Y\left(\left\lfloor \frac{r_j}{k} \right\rfloor\right) + \bar{F}_Y\left(\left\lfloor \frac{r_i}{k} \right\rfloor\right)\right) \leq T,$$

which implies

$$M \geq \frac{1}{T}\sum_{(i,j)\in\mathbb{Z}\times\mathbb{Z},1\leq i,j\leq M,i\neq j} \left(\bar{F}_Y\left(\left\lfloor \frac{r_j}{k} \right\rfloor\right) + \bar{F}_Y\left(\left\lfloor \frac{r_i}{k} \right\rfloor\right)\right). \tag{2.6}$$

Using lemma 4 we get

$$\begin{aligned} M &\geq \frac{1}{T}\sum_{(i,j)\in\mathbb{Z}\times\mathbb{Z},1\leq i,j\leq M,i\neq j} \left(2\sum_{y=\left\lceil \frac{r-\delta}{k}\right\rceil}^{\lfloor r-\delta\rfloor} \frac{\binom{r-\delta}{y}\binom{n-r-\delta}{r-\delta-y}}{\binom{n}{r+\delta}}\right) \\ &= \frac{1}{T}\frac{M(M-1)}{2}\left(2\sum_{y=\left\lceil \frac{r-\delta}{k}\right\rceil}^{\lfloor r-\delta\rfloor} \frac{\binom{r-\delta}{y}\binom{n-r-\delta}{r-\delta-y}}{\binom{n}{r+\delta}}\right), \end{aligned} \tag{2.7}$$

which implies

$$M \leq \frac{T}{\sum_{y=\lceil \frac{r-\delta}{k} \rceil}^{\lfloor r-\delta \rfloor} \frac{\binom{r-\delta}{y}\binom{n-r-\delta}{r-\delta-y}}{\binom{n}{r+\delta}}} + 1. \tag{2.8}$$

The expected (r, T, k, δ)-subset capacity of V should be bounded above by this expression and the tightness of the bound will depend on the parameter δ. □

2.3 Empirical Results

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.