Expected Subset Capacity (Draft)

Chandradeep Chowdhury
Patrick Perrine

November 27, 2023

In this article, we explore the notion of *interference* of randomly shared items between sets within a universe. We define and generalize our probabilistic framework as a result of the hypergeometric distribution, and then arrive at results regarding the *capacity* of the universe based on those notions of interference.

TODO: Confirm if this simplification to work with general sets is valid.

1 Basic Interference Framework

Definition 1. (**Interference**) Given two sets U, W, we say U *interferes* with W if

$$|U \cap W| \geq |W|. \tag{1}$$

Lemma 2. *Given a set V with n items and two proper subsets U, W of respective sizes r_u, r_w, the probability of an interference between U and W, denoted as Y, follows a hypergeometric distribution with $Y \sim Hypergeometric(n, r_u, y)$.*

Proof. If $V = \{v_1, ..., v_n\}$, we can represent a 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, W intersect at the indices where both Boolean vectors u, w have a 1. Let Y be a discrete random variable denoting the number of indices where both u, w have a 1. Then

$$\mathbb{P}(Y = y) = \frac{\binom{r_u}{y}\binom{n-r_u}{r_w-y}}{\binom{n}{r_w}}. \tag{2}$$

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 $\binom{n}{r_w}$. For brevity, we then use Vandermonde's identity to realize this probability as a result of the hypergeometric distribution, with $Y \sim Hypergeometric(n, r_u, y)$. □

2 Generalizing Interference for k Instances

Definition 3. (***k*-Interference**) Given two sets U, W, and for some $k \in (0, |W|]$, we say U *$k-$interferes* with W if

$$|U \cap W| \geq \frac{|W|}{k}. \tag{3}$$

Corollary 4. *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. If $\frac{|W|}{k} = 1$, we could arrive at a variation of the Hitting Set problem.

Lemma 5. *Given a set V with n items and two proper subsets U, W of respective sizes r_u, r_w, the probability that U k-interferes with W, denoted by Y, is the tail distribution of Y at $\lfloor \frac{r_w}{k} \rfloor$, or simply $\bar{F}_Y\left(\lfloor \frac{r_w}{k} \rfloor\right)$.*

Proof. We know that the minimum number of k-interferences possible is $\lceil \frac{r_w}{k} \rceil$. We can then, for brevity, show the probability that U k-interferes with W is simply

$$\sum_{y=\lceil \frac{r_w}{k} \rceil}^{r_w} \mathbb{P}(Y = y) = \mathbb{P}\left(Y \geq \left\lceil \frac{r_w}{k} \right\rceil\right) = \mathbb{P}\left(Y > \left\lfloor \frac{r_w}{k} \right\rfloor\right) = \bar{F}_Y\left(\left\lfloor \frac{r_w}{k} \right\rfloor\right). \tag{4}$$

□

TODO: Confirm whether we can simply flip the ceiling function to a floor function here, or that we need to subtract 1 from the ceiling function here. Recall that $\bar{F}_X(x) = \mathbb{P}(X > x)$ and that $\mathbb{P}(Y \geq y) \neq \mathbb{P}(Y > y)$ because Y is discrete.

Lemma 6. *If*

1. $\mathbb{E}[r_u] = \mathbb{E}[r_w] = r$,
2. $r_u, r_w \in [r - \delta, r + \delta]$ *for some* $\delta \in \mathbb{Z}$,
3. $n > kr^2/r_w$,
4. $Y \sim Hypergeometric(n, r_u, y)$,

then we observe that

$$\mathbb{E}\left[\bar{F}_Y\left(\left\lfloor \frac{r_w}{k} \right\rfloor\right)\right] \geq \sum_{y=\lceil \frac{r-\delta}{k} \rceil}^{r-\delta} \left(\frac{r^2 - ry}{ny - ry}\right)^y \left(\frac{nr - r^2}{nr - ny}\right)^{r-\delta}.$$

Proof. Given our assumptions, we observe that

$$\begin{aligned}
\mathbb{E}\left[\bar{F}_Y\left(\left\lfloor \frac{r_w}{k} \right\rfloor\right)\right] &= \mathbb{E}\left[\sum_{y=\lceil \frac{r_w}{k} \rceil}^{r_w} \mathbb{P}(Y = y)\right] \\
&= \mathbb{E}\left[\sum_{y=\lceil \frac{r_w}{k} \rceil}^{r_w} \frac{\binom{r_u}{y}\binom{n-r_u}{r_w-y}}{\binom{n}{r_w}}\right] \\
&= \sum_{y=\lceil \frac{r_w}{k} \rceil}^{r_w} \mathbb{E}\left[\frac{\binom{r_u}{y}\binom{n-r_u}{r_w-y}}{\binom{n}{r_w}}\right] \\
&\geq \sum_{y=\lceil \frac{r_w}{k} \rceil}^{r_w} \mathbb{E}\left[\frac{\left(\frac{r_u}{y}\right)^y \left(\frac{n-r_u}{r_w-y}\right)^{r_w-y}}{\left(\frac{n}{r_w}\right)^{r_w}}\right] \\
&= \sum_{y=\lceil \frac{r_w}{k} \rceil}^{r_w} \frac{\left(\frac{r}{y}\right)^y \left(\frac{n-r}{r-y}\right)^{r_w-y}}{\left(\frac{n}{r}\right)^{r_w}} \\
&= \sum_{y=\lceil \frac{r_w}{k} \rceil}^{r_w} \left(\frac{r^2 - ry}{ny - ry}\right)^y \left(\frac{nr - r^2}{nr - ny}\right)^{r_w} \\
&\geq \sum_{y=\lceil \frac{r-\delta}{k} \rceil}^{r-\delta} \left(\frac{r^2 - ry}{ny - ry}\right)^y \left(\frac{nr - r^2}{nr - ny}\right)^{r-\delta}.
\end{aligned} \tag{5}$$

□

3 Capacity Results

Definition 7. (**(r,T,k)-Subset Capacity**) Given a set $V = \{v_1, ..., v_n\}$, the (r, T, k)-*subset capacity* of V is the *maximum* number of subsets of expected size r that can be collected subject to the constraint that for any randomly picked subset U,

$$\mathbb{E}[X] \leq T \tag{6}$$

where X is the number of interferences caused due to picking U.

3.1 Capacity for Exactly r-sized Subsets

Theorem 8. *Given a set V with n items and the property that every subset will have size exactly r, the (r, T, k)-subset capacity of V is*

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

Proof. Suppose we have M subsets in the universe. Pick an arbitrary subset U. From lemma 5, we know that the probability of U k-interfering with another arbitrary subset W is $\bar{F}_Y\left(\left\lfloor \frac{r_w}{k} \right\rfloor\right)$. Since there are $M-1$ other subsets and we know $r_w = r$, the 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 6, we have

$$(M-1)\bar{F}_Y\left(\left\lfloor \frac{r}{k} \right\rfloor\right) \leq T \implies M \leq \frac{T}{\bar{F}_Y\left(\left\lfloor \frac{r}{k} \right\rfloor\right)} + 1. \tag{7}$$

The (r, T, k)-subset capacity of V is the largest integer M that satisfies inequality 7. □

Alternate proof. Suppose we have M subsets in the universe. Pick two subsets U, W without replacement. From lemma 5, we know that the probability of U k-interfering with another arbitrary subset W is $\bar{F}_Y\left(\left\lfloor \frac{r_w}{k} \right\rfloor\right)$. Since we know all subsets have the same size, by corollary 4 this becomes the probability that U, W pair will cause exactly 2 k-interferences. 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 pairings 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 6, we have

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

The (r, T, k)-subset capacity of V is the largest integer M that satisfies inequality 7. □

3.2 Capacity for Expected r-sized Subsets

Definition 9. (**Expected (r, T, k, δ)-Subset Capacity**) Given a set $V = \{v_1, ..., v_n\}$, the *expected (r, T, k, δ)-subset capacity* of V is the *maximum* number of subsets that can be picked subject to the conditions that for any randomly picked subset U with size r', such that

1. $\mathbb{E}[r'] = r$,
2. $r' \in [r - \delta, r + \delta]$,
3. $n > kr^2/r'$,
4. $E[X] \leq T$ where X is the number of interferences caused due to picking U.

Theorem 10. *Given a set V with n items, the expected (r, T, k, δ)-subset capacity of V is*

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

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 5, 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:

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

TODO: In trying to generalize this to work with general sets, is this sum neccessary? Or does it just need to be changed?

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 4, 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

$$M\geq\frac{1}{T}\sum_{(i,j)\in\mathbb{Z}\times\mathbb{Z},1\leq i,j\leq M,i\neq j}\left(\bar{F}_Y\left(\left\lfloor\frac{r_j}{k}\right\rfloor\right)+\bar{F}_Y\left(\left\lfloor\frac{r_i}{k}\right\rfloor\right)\right). \tag{8}$$

Taking the expectation on both sides and using lemma 6 we get

$$\begin{aligned}
M=\mathbb{E}[M]&\geq\mathbb{E}\left[\frac{1}{T}\sum_{(i,j)\in\mathbb{Z}\times\mathbb{Z},1\leq i,j\leq M,i\neq j}\left(\bar{F}_Y\left(\left\lfloor\frac{r_j}{k}\right\rfloor\right)+\bar{F}_Y\left(\left\lfloor\frac{r_i}{k}\right\rfloor\right)\right)\right]\\
&=\frac{1}{T}\sum_{(i,j)\in\mathbb{Z}\times\mathbb{Z},1\leq i,j\leq M,i\neq j}\left(\mathbb{E}\left[\bar{F}_Y\left(\left\lfloor\frac{r_j}{k}\right\rfloor\right)\right]+\mathbb{E}\left[\bar{F}_Y\left(\left\lfloor\frac{r_i}{k}\right\rfloor\right)\right]\right)\\
&\geq\frac{1}{T}\sum_{(i,j)\in\mathbb{Z}\times\mathbb{Z},1\leq i,j\leq M,i\neq j}\left(2\cdot\sum_{y=\left\lceil\frac{r-\delta}{k}\right\rceil}^{r-\delta}\left(\frac{r^2-ry}{ny-ry}\right)^y\left(\frac{nr-r^2}{nr-ny}\right)^{r-\delta}\right)\\
&\stackrel{?}{=}\frac{1}{T}\frac{M(M-1)}{2}\left(2\cdot\mathbb{E}\left[\bar{F}_Y\left(\left\lfloor\frac{r}{k}\right\rfloor\right)\right]\right),
\end{aligned} \tag{9}$$

(**TODO:** Confirm if the final statement in equation 9 is valid.)

which implies

$$M\leq\frac{T}{\mathbb{E}\left[\bar{F}_Y\left(\left\lfloor\frac{r}{k}\right\rfloor\right)\right]}+1. \tag{10}$$

The expected (r,T,k,δ)-subset capacity of V is the largest integer that satisfies inequality 10.

□

4 Conclusion

We happened upon these set-theoretic results when working with induced subgraphs intended for knowledge representation. One example of an application of our work could be that of the Neuroidal Model by Leslie Valiant, in which induced subgraphs are used as memories of a neural system within a computational neuroscience perspective.