Systems of Overlap Representation for Families of Intervals

Adam S. Jobson*, André E. Kézdy†, Kevin G. Milans‡,
Douglas B. West§, D. Jacob Wildstrom¶

July 20, 2021

Abstract

Two sets *overlap* if they intersect and neither contains the other. Given a family $\mathcal{F}$ of sets, a *system of overlap representation (SOR)* for $\mathcal{F}$ assigns to each set $X \in \mathcal{F}$ a subset S_X of X, called its *representative set*, such that the representative sets chosen for any two overlapping members of $\mathcal{F}$ intersect. Let $\mathcal{F}_n$ be the family of intervals of integers contained in $\{1, \ldots, n\}$, and let $f(n)$ be the minimum of the maximum size of the sets in an SOR of $\mathcal{F}_n$. We prove $(1 - o(1)) \lg(n-1) < f(n) \le 2 \lg(n-1)$.

MSC Codes: 05D05, 05D15

1 Introduction

Let $\mathcal{F}$ be a family $A_1, \ldots, A_m$ of sets. A *system of distinct representatives (SDR)* for $\mathcal{F}$ is a set $\{z_1, \ldots, z_m\}$ of distinct elements such that $z_i \in A_i$ for $1 \le i \le m$. An obvious necessary condition for the existence of an SDR is *Hall's Condition*, requiring $\left|\bigcup_{i \in S} A_i\right| \ge |S|$ for all $S \subseteq [m]$. Hall [2] proved that this condition is also sufficient.

In an SDR, each set is trimmed down to a single representative. In this note, we want to reduce the sets in a family to representing subsets in a way that captures an aspect of interaction between the sets. Two sets *overlap* if they intersect and neither contains the other. A *system of overlap representation (SOR)* for a family $\mathcal{F}$ of sets assigns to each set X in $\mathcal{F}$ a *representative set* S_X contained in X so that S_X and S_Y have nonempty intersection whenever X and Y overlap.

*Department of Mathematics, University of Louisville: adam.jobson@louisville.edu.

†Department of Mathematics, University of Louisville: andre.kezdy@louisville.edu.

‡Department of Mathematics, West Virginia University: kmilans@gmail.com.

§Departments of Mathematics, Zhejiang Normal University, Jinhua, China, and University of Illinois, Urbana, IL: west@math.uiuc.edu. Research supported by National Natural Science Foundation of China grants NSFC 11871439, 11971439, and U20A2068.

¶Department of Mathematics, University of Louisville: david.wildstrom@louisville.edu.

The sets in an SDR are small, just size 1. Similarly, we seek SORs with small representative sets. This could be useful when sets are expensive to access and we need to study the interactions between one set in the family and the sets with which it overlaps by treating common members. It should be noted that an SOR does not encode the overlap relation, in the sense that S_X and S_Y may also intersect when $X \subseteq Y$.

Motivation for studying SORs comes from the data visualization community. In order to design efficient representations of set-based data arising in application areas, this community has studied modified Euler diagrams (see, for example, [1, 5]). As described in [5], the term "Euler diagram" is used for "graphical representations that depict sets and their intersections. In an Euler diagram, a set is a region of the plane bounded by a curve and intersections between sets are depicted through overlaps between these regions".

Systems of overlap representation offer a compact alternative to visualize families of intervals where containments are forbidden. Such diagrams may be particularly valuable to display, preview, or survey large catalogues of interval-based data using only a small number of feature-rich samples. Such data can arise from text, time-series, electrocardiograms, or other settings involving location along a single dimension. Algorithms seeking pertinent data often sample data sets looking for exemplars (particular data points exhibiting desirable features). For example, in [3] the problem of detecting anomalies in time-series data in various applications is considered. Carefully arranged exemplars would speed such searches. An SOR that lists exemplars using a logarithmic number of samples could efficiently guide such searches. Data visualization becomes more important as data sets grow and machines must communicate with humans about them.

When $\mathcal{F}' \subseteq \mathcal{F}$ and $\mathcal{G}$ is an SOR of $\mathcal{F}$, the family $\mathcal{G}'$ consisting of $\{S_X \in \mathcal{G} \colon X \in \mathcal{F}'\}$ is an SOR of $\mathcal{F}'$. Therefore, to understand how big the sets in an SOR might need to be, it makes sense to study the extremal problem when $\mathcal{F}$ has many sets on a fixed ground set. We consider families of integer intervals.

Given integers a and b with $a \leq b$, let $[a, b]$ denote $\{i \in \mathbb{Z} \colon a \leq i \leq b\}$. Let $\mathcal{F}_n$ denote the family of intervals $[a, b]$ such that $1 \leq a \leq b \leq n$. Let $f(n)$ be the minimum of the maximum size of the sets in an SOR of $\mathcal{F}_n$. We determine $f(n)$ asymptotically within a factor of 2, proving

$$(1 - o(1)) \lg n \leq f(n) \leq 2 \lg(n - 1),$$

where lg denotes the base-2 logarithm.

The problem of determining $f(n)$ can be expressed as an integer linear program. First we express it using a quadratic program. Introduce variables $x_{I,r}$ for all $r \in I \in \mathcal{F}_n$, with $x_{I,r}$ being the 0, 1-indicator variable for $r \in S_I$. With $x_{I,r} \in \{0, 1\}$, we then seek to minimize z such that $\sum_{r \in I \cap J} x_{I,r} x_{J,r} \geq 1$ when I and J overlap and $\sum_{r \in I} x_{I,r} \leq z$ for $I \in \mathcal{F}_n$.

The quadratic program can be converted to a linear program by a standard trick. For each $r \in I \cap J$ with I and J overlapping, replace $x_{I,r} x_{J,r}$ with an auxiliary 0, 1-variable $w_{I,J,r}$ satisfying $w_{I,J,r} \leq x_{I,r}$, $w_{I,J,r} \leq x_{J,r}$, and $x_{I,r} + x_{J,r} \leq 1 + w_{I,J,r}$. The linear program has confirmed $f(n) = \lceil \lg(n - 1) \rceil$ for $2 \leq n \leq 18$, but the asymptotic behavior remains open.

Question 1. What is the asymptotic behavior of $f(n)$?

2 Powers of 2 and the Upper Bound

Our proof of the upper bound on $f(n)$ uses a number-theoretic property that has other applications, which we will mention after the proof.

Theorem 1. *For n a positive integer, $f(n) \leq 2\lg(n-1)$.*

Proof. Consider an interval $[a, b]$ in $\mathcal{F}_n$. We specify a representative set $S_{a,b}$ for the interval $[a, b]$. For each k with $0 \leq k \leq \lfloor \lg(n-1) \rfloor$ such that 2^k divides some number in $[a, b]$, we include in $S_{a,b}$ the smallest and largest numbers in $[a, b]$ that are divisible by 2^k. Note that when $b = n$, the values n is in $S_{a,b}$ due to $k = 0$, so allowing $k = \lg n$ would be redundant. When $k = \lfloor \lg(n-1) \rfloor$ there can only be one value contributed by k, so $|S_{a,b}| \leq 1 + 2\lg(n-1)$. We first prove that this construction suffices, and then we show that the contribution to $S_{a,b}$ using $k = \lfloor \lg(n-1) \rfloor$ is not needed.

Given overlapping intervals $[a, b]$ and $[c, d]$ with $a \leq c \leq b \leq d$, let x be the largest power of 2 that divides some integer in $[c, b]$. If two consecutive multiples of x lie in $[c, b]$, then one of them is an even multiple of x, divisible by $2x$, which is a larger power of 2. Thus x divides exactly one number r in $[c, b]$. This makes r the least multiple of x in $[c, d]$ and the greatest multiple of x in $[a, b]$. Hence $r \in S_{a,b} \cap S_{c,d}$, and the representative sets form an SOR.

Finally, let $k = \lfloor \lg(n-1) \rfloor$, and suppose that 2^k is the only element of $[a, b] \cap [c, d]$. If there is no instance of this, then we can discard the contribution due to k from each representative set. If 2^k is not contributed to the intersection by 2^{k-1}, then one of the intervals contains $[2^{k-1}, 3 \cdot 2^{k-1}]$. Since the intervals overlap, the other interval also contains 2^{k-1} or $3 \cdot 2^{k-1}$ (since it contains 2^k) Now, since $4 \cdot 2^{k-1} > n$, that value lies in $S_{a,b} \cap S_{c,d}$. □

Note that the representative set for an interval $[a, b]$ given in the proof of Theorem 1 is the same for all n with $n \geq b$. This may explain why the upper bound is not sharper.

In the proof, we used the fact that the largest power of 2 that divides some integer in an interval $[c, b]$ divides only one number in that interval. This property also yields a short proof that a nontrivial difference between two harmonic numbers cannot be an integer. The *harmonic number* H_n is $\sum_{i=1}^{n} 1/i$. Theisinger [6] proved that H_n is not an integer when $n > 1$, and Kürschák [4] gave a proof using the property stated here. The same technique proves the claim more generally for differences of harmonic numbers.

If the difference $H_n - H_m$ is an integer k, then $\sum_{i=m+1}^{n} n!/i = n!k$. Let 2^s be the highest power of 2 dividing a number in the interval $[m+1, n]$, and let 2^t be the highest power of 2 dividing $n!$. Since 2^s divides only one number in the interval, 2^{t-s+1} divides all but one term on the left, and hence it does not divide $n!k$. Since 2^t does divide $n!k$, we conclude $s = 0$, and hence $m = n - 1$. The interval is a single odd integer, so $n = 1$.

3 The Lower Bound

We prove a lower bound on $f(n)$ in terms of a parameter s. The special case $s = 1$ yields a simple argument for a lower bound of $(1/2)\lg n$, but setting $s = \lfloor \lg n \rfloor$ improves the lower bound by an asymptotic factor of 2.

Theorem 2. *For n a positive integer, $f(n) > (1 - o(1)) \lg n$.*

Proof. Due to the monotonicity of f, it suffices to prove this lower bound for infinitely many values of n. We will let $s = \lfloor \lg n \rfloor$ and assume that n is a multiple of s (essentially, we are proving the lower bound by considering intervals up to the greatest multiple of s bounded by n). Let $m = n/s$.

Partition $[n]$ into s subintervals of length m; call them $X_1, \ldots, X_s$. (That is, $X_i = [(i-1)m + 1, im]$ for $1 \le i \le s$.) For $1 \le i \le s-1$, let A_i be the family of intervals contained in $X_i \cup X_{i+1}$ that have the same number of elements from X_i and X_{i+1}. (That is, $A_i = \{[im - j, im + 1 + j] \colon 0 \le j \le m - 1\}$, with left endpoint in X_i and right endpoint in X_{i+1}.) Let A_0 consist of the subintervals of X_1 that contain the element 1, and let A_s consist of the subintervals of X_s that contain the element n. Note that $|A_i| = m$ for $0 \le i \le s$.

Let $\mathcal{F} = \bigcup_{i=0}^{s} A_i$, so $|\mathcal{F}| = (s+1)m$. It suffices to show that every SOR for $\mathcal{F}$ contains some representative set of size at least $\frac{s}{s+1} \lg m$. Since $\frac{s}{s+1} \ge -1 + \lg n \lg n$, this yields $f(n) \ge (1 - 1/\lg n)(\lg n - \lg\lg n)$. Given an SOR for $\mathcal{F}$, for $1 \le i \le s$ let $T_i = \{(I, r) \colon I \in A_{i-1} \cup A_i \text{ and } r \in X_i \cap S_I\}$. We obtain a lower bound on $|T_i|$ and then divide by $|\mathcal{F}|$ to obtain a lower bound on $f(n)$.

Call the elements 1 and n the "trivial" endpoints. Given $(I, r) \in T_i$, there is a unique nontrivial endpoint of I contained in X_i; call this value p when i is specified. Group the members (I, r) of T_i according to the logarithm of $|r - p|$. In particular, for $j \ge 1$ let $T_{i,j}$ be the subset of T_i consisting of all pairs (I, r) such that $2^{j-1} \le |r - p| \le 2^j - 1$. Let $T_{i,0}$ be the set of pairs $(I, r) \in T_i$ such that r itself is the nontrivial endpoint of I in X_i.

For $I \in A_i$ with $i \ge 1$, there is an interval $I' \in A_{i-1}$ that intersects I precisely at its lower endpoint. Similarly, when $i < s$, some interval I' in A_{i+1} intersects I precisely at its upper endpoint. The intervals I and I' overlap (if neither is $\{1\}$ or $\{n\}$). Thus every interval in A_i must have its nontrivial endpoints in its representative set. Each element of $[n] - \{1, n\}$ is a nontrivial endpoint for two intervals in $\mathcal{F}$, and 1 and n are never nontrivial endpoints. Hence $|\bigcup_{i=1}^{s} T_{i,0}| = 2(n-2)$.

Let $k = \lfloor \lg m \rfloor$. We claim $|T_{i,j}| \ge m - 2^j - 1$ for $1 \le j \le k$. Since $|X_i| = m$, there are $m - 2^j - 1$ subintervals of X_i having size 2^j that are disjoint from $\{\min X_i, \max X_i\}$. For each such subinterval $[a, b]$, we next obtain a member of $T_{i,j}$. The interval $I \in A_{i-1}$ ending at b and the interval $I' \in A_i$ starting at a overlap and have intersection $[a, b]$. Hence some point $r \in [a, b]$ is in both S_I and $S_{I'}$. The interval $[a, b]$ has 2^j elements. If r is in the lower half, then its distance from b is at least 2^{j-1} and is less than 2^j, putting (I, r) into $T_{i,j}$. Otherwise, the same reasoning puts (I', r) into $T_{i,j}$. This completes the proof of the lower bound on $|T_{i,j}|$.

Now we compute

$$\begin{aligned}\left|\bigcup_{i=1}^{s} T_i\right| \geq \sum_{i=1}^{s}\sum_{j=0}^{k} |T_{i,j}| &\geq 2(n-2) + s\sum_{j=1}^{k}\left(m - 2^j - 1\right) \\ &= 2(n-2) + ks(m-1) - 2s(2^k - 1) \\ &\geq (2s + ks - 2s)(m-1) \\ &= ks(m-1).\end{aligned}$$

By the pigeonhole principle, some interval I in $\mathcal{F}$ has a representative set of size at least $|T|/|\mathcal{F}|$, where $T = \bigcup_{i=1}^{s} T_i$. Since $|\mathcal{F}| = (s+1)m = (s+1)n/s$,

$$\frac{|T|}{|\mathcal{F}|} \geq \frac{s}{s+1} \cdot k \cdot \frac{n-s}{n} \approx \left(1 - \frac{1}{\lg n}\right) \lfloor \lg n - \lg\lg n \rfloor \left(1 - \frac{\lg n}{n}\right) \sim \lg n. \qquad \square$$

One can also seek higher-dimensional generalizations involving overlaps in the family of rectangles $[a, b] \times [c, d]$ with $1 \leq a \leq b \leq n$ and $1 \leq c \leq d \leq n$.

Acknowledgment

We thank Keith Conrad for providing the references for the nonintegrality of harmonic numbers.

References

[1] J. Burton and G. Stapleton, Special issue on Euler and Venn diagrams: Guest editors' introduction, *J. Log. Lang. Inf.* **24** (2015), 357–359.

[2] P. Hall, On representatives of subsets, *J. Lond. Math. Soc.* **10** (1935), 26–30.

[3] M. Jones, D. Nikovski, M. Imamura, and T. Hirata, Exemplar learning for extremely efficient anomaly detection in real-valued time series, *Data Min. Knowl. Discov.* **30** (2016), 1427–1454.

[4] J. Kürschák, A Harmonikus Sorról, *Mat. és Fiz. Lapok* 27 (1918), 299–300.

[5] P. Simonetto and D. Archambault, Fully automatic visualisation of overlapping sets, *Computer Graphics Forum* **28** (2009), 967–974.

[6] L. Theisinger, Bemerkung über die harmonische Reihe, *Monatsh. f. Mathematik und Physik* 26 (1915), 132–134.