A question of Erdős and Graham on Egyptian fractions

David Conlon* Jacob Fox† Xiaoyu He‡ Dhruv Mubayi§ Huy Tuan Pham¶
Andrew Suk‖ Jacques Verstraëte**

Abstract

Answering a question of Erdős and Graham, we show that for each fixed positive rational number x the number of ways to write x as a sum of reciprocals of distinct positive integers each at most n is $2^{(c_x+o(1))n}$ for an explicit constant c_x increasing with x.

1 Introduction

The study of Egyptian fractions, that is, sums of reciprocals of distinct positive integers, has a long history in combinatorial number theory (see, for example, [4]). The fact that every positive fraction can be written as an Egyptian fraction goes back at least to work of Fibonacci at the start of the 13th century. Much more recently, a result of Bloom [2] says that any subset of the natural numbers of positive upper density has a finite subset the sum of whose reciprocals adds to one.

In this paper, we will be concerned with a problem raised by Erdős and Graham [8, Page 36] in 1980 (see also [3, Problem 297]): how many ways are there to write one as a sum of distinct unit fractions with denominator at most n? Very recently, Steinerberger [11] showed that the number of such Egyptian fractions is at most $2^{0.93n}$. This already answered one particular question of Erdős and Graham, who asked whether the answer was $2^{n-o(n)}$. Here we answer their question much more precisely by showing that the count is $2^{(1+o_n(1))cn}$ for an explicit constant $c \approx 0.91117$.

Our main result more generally estimates the number of Egyptian fractions summing to any fixed positive rational. Let $h : [0,1] \to \mathbb{R}$ be given by $h(p) = -p\log_2 p - (1-p)\log_2(1-p)$ for $p \in (0,1)$ and $h(0) = h(1) = 0$.

Theorem 1. *For any fixed $x \in \mathbb{Q}_{>0}$, the number of subsets $A \subseteq [n]$ with $x = \sum_{a\in A} 1/a$ is $2^{c_x n+o(n)}$, where*

$$c_x := \int_0^1 h\left(\frac{1}{1+e^{\lambda/y}}\right) dy$$

and λ is the unique real number such that

$$\int_0^1 \frac{1}{y(1+e^{\lambda/y})} dy = x.$$

In particular, c_x is a strictly increasing function with $c_0 = 0$, $c_1 \approx 0.91117$ and $c_x \to 1$ as $x \to \infty$.

Our proof has two main steps. In the first step, we use entropy methods to show that $2^{(c_x+o(1))n}$ is the correct asymptotic count for the number of Egyptian fractions formed by adding distinct unit fractions with denominator at most n whose sum is at most x. Then, in the second step, we use a method reminiscent of the absorption technique in extremal graph theory to show that the same asymptotic count holds for the number of Egyptian fractions summing to exactly x. Very roughly, we first set aside a small reservoir subset of $[n]$. Then, after finding many subsets of $[n]$ disjoint from this reservoir whose sums of reciprocals are somewhat smaller than x and whose denominators have no very large prime power factors, we iteratively 'clean' these fractions by adding unit fractions from the reservoir to obtain a sum $x' < x$ with small denominator. This is accomplished through the use of a recent result [5] on the existence of homogeneous generalized arithmetic progressions in subset sums. Finally, we find a small subset of the reservoir whose sum of reciprocals is equal to the remaining difference $x - x'$.

*Department of Mathematics, California Institute of Technology, Pasadena, CA 91125. Email: dconlon@caltech.edu.
†Department of Mathematics, Stanford University, Stanford, CA 94305. Email: jacobfox@stanford.edu.
‡Department of Mathematics, Princeton University, Princeton, NJ 08544. Email: xiaoyuh@princeton.edu.
§Department of Mathematics, Statistics and Computer Science, University of Illinois, Chicago, IL 60607. Email: mubayi@uic.edu.
¶Department of Mathematics, Stanford University, Stanford, CA 94305. Email: huypham@stanford.edu.
‖Department of Mathematics, University of California at San Diego, La Jolla, CA 92093. Email: asuk@ucsd.edu.
**Department of Mathematics, University of California at San Diego, La Jolla, CA 92093. Email: jacques@ucsd.edu.

2 Counting through entropy

In this section, we will use entropy methods to estimate the number of subsets $A \subseteq [n]$ with $s(A) := \sum_{a \in A} 1/a \leq x$. To state our result, we need some notation. As in the introduction, let $h(p) = -p \log_2 p - (1-p)\log_2(1-p)$ for $0 < p < 1$ and $h(0) = h(1) = 0$. Given $x > 0$, we choose $p_1, \dots, p_n \in [0,1]$ so as to maximize $\sum_{m=1}^n h(p_m)$ given $\sum_{m=1}^n p_m/m \leq x$. We then let $P(x)$ be the distribution of $(Y_1, \dots, Y_n)$, where each $Y_m = 1$ with probability p_m and 0 otherwise independently of each other. Then the (Shannon) entropy of $P(x)$ is given by $H(P(x)) = \sum_{m=1}^n h(p_m)$.

Lemma 2. *For $\varepsilon > 0$ and $x \in (0, (1-\varepsilon)(\ln n)/2)$, the number of subsets $A \subseteq [n]$ with $s(A) \leq x$ is bounded above by $2^{H(P(x))}$. Furthermore, there is $c_{x,n} > 0$ with $c_{x,n} = \Theta(e^{-2x})$ such that, for any $U \subseteq [n]$, the number of subsets $A \subseteq U$ with $s(A) \leq x$ is bounded below by $2^{H(P(x)) - (n - |U|) - O(\sqrt{n/c_{x,n}})}$.*

We first characterize $P(x)$.

Lemma 3. *For $x < (1/2)(\sum_{m=1}^n 1/m)$, the distribution $P(x)$ is given by setting $p_m = \frac{1}{1+e^{c_{x,n} n/m}}$ for the unique $c_{x,n}$ such that $\sum_{m=1}^n p_m/m = x$. Moreover, there is $C > 0$ such that, provided $x < (1-\varepsilon)(\ln n)/2$, $c_{x,n} > 0$ and $c_{x,n} \in [C^{-1}e^{-2x}, Ce^{-2x}]$. Finally, for $\varepsilon > 0$ and assuming $x \in [\varepsilon, 1/\varepsilon]$, $c_{x,n} = \lambda + o_n(1)$, where $\lambda > 0$ is the unique solution to*

$$\int_0^1 \frac{1}{y(1+e^{\lambda/y})} dy = x.$$

In particular,

$$H(P(x)) = (1+o_n(1))n \int_0^1 h\left(\frac{1}{1+e^{\lambda/y}}\right) dy = \Theta(n). \tag{1}$$

Note that here the $o_n(1)$ terms may depend on ε.

Proof. Note that $H(P(x))$ is strongly convex and bounded (as a function of $p_1, \dots, p_n$) in $[0,1]^n$. The unique stationary point of $H(P(x))$ in $[0,1]^n$ is $p_1 = \cdots = p_n = 1/2$. Thus, for $x < (1/2)(\sum_{m=1}^n 1/m)$, the maxima of $H(P(x))$ must be achieved on the boundary of $[0,1]^n \cap \{\sum_m p_m/m \leq x\}$ and, since $h(x) = 0$ for $x \in \{0,1\}$, we must have that the maxima are achieved on $\sum_m p_m/m = x$. At a maximum $(p_1^*, \dots, p_n^*)$, we must have that $\nabla H(P(x))$ is parallel to $(1, 1/2, \dots, 1/n)$. Noting that $h'(p) = \log \frac{1-p}{p}$, we obtain that such a point satisfies $p_m^* = \frac{1}{1+e^{c'/m}}$ for some c'. By the condition $\sum_{m \leq n} p_m^*/m = x$, we must also have that c' satisfies

$$\sum_{m=1}^n \frac{1}{m(1+e^{c'/m})} = x.$$

Letting $c = c_{x,n} = c'/n$, we have that

$$\frac{1}{n} \sum_{m=1}^n \frac{1}{(m/n)(1+e^{c/(m/n)})} = x.$$

It is easy to check that $c > 0$ when $x < (1-\varepsilon)(\ln n)/2$ and, for x sufficiently large, that $c = \Theta(e^{-2x})$. Indeed, for the last estimate, we observe that

$$\sum_{m=cn/(k+1)}^{cn/k} \frac{1}{m(1+e^{cn/m})} = \Theta(e^{-k}k^{-1}),$$

so

$$\sum_{m \leq cn} \frac{1}{m(1+e^{cn/m})} = O(1).$$

For $1 \leq k \leq 1/c$,

$$\sum_{m=cnk}^{cn(k+1)} \frac{1}{m(1+e^{cn/m})} \leq \frac{1}{k(1+e^{1/(k+1)})}, \qquad \sum_{m=cnk}^{cn(k+1)} \frac{1}{m(1+e^{cn/m})} \geq \frac{1}{(k+1)(1+e^{1/k})}.$$

Note that $1 + e^{1/k} = 2 + O(1/k)$, so we obtain that

$$\sum_{k=1}^{1/c} \frac{1}{k(1+e^{1/(k+1)})}, \sum_{k=1}^{1/c} \frac{1}{(k+1)(1+e^{1/k})} = \frac{1}{2}\ln(1/c) + O(1),$$

from which we immediately deduce the desired estimate on c.

Finally, for $x \in [\varepsilon, 1/\varepsilon]$ and $c > 0$, we can approximate $\frac{1}{n}\sum_{m=1}^{n} \frac{1}{(m/n)(1+e^{c/(m/n)})}$ by the integral $\int_0^1 \frac{1}{y(1+e^{c/y})}dy$. We thus obtain that for $\lambda > 0$ satisfying $\int_0^1 \frac{1}{y(1+e^{\lambda/y})}dy = x$, we have $c_{x,n} = \lambda + o_n(1)$, from which (1) follows readily. □

We will use the following version of the standard Berry–Esseen bound [1, 7, 10] in the proof of Lemma 2.

Lemma 4. *Let $X_1, \dots, X_n$ be independent centered random variables with $\mathbb{E}[X_i^2] = \zeta_i$ and $\mathbb{E}[X_i^3] = \rho_i$. Let $Z = \frac{\sum_{i=1}^n X_i}{(\sum_{i=1}^n \zeta_i)^{1/2}}$ and Z' be a standard Gaussian. Then*

$$\sup_{y\in\mathbb{R}} |\Pr[Z \le y] - \Pr[Z' \le y]| \le C \frac{\sum_{i=1}^n \rho_i}{(\sum_{i=1}^n \zeta_i)^{3/2}}.$$

Proof of Lemma 2. Let S be a finite set of real numbers and $n = |S|$. Let $r_S(x)$ be the number of subsets of S that sum to at most x. Define the random variable X to be a uniform random subset of S whose elements sum to at most x. Note that the entropy of X satisfies $H(X) = \log_2 r_S(x)$, so $r_S(x) = 2^{H(X)}$. For each $s \in S$, let X_s be the indicator random variable of the event $s \in X$ and let $p_s = \Pr[X_s]$, so that $\mathbb{E}[\sum_{s\in S} sX_s] = \sum_{s\in S} sp_s \le x$. Observe that X has the same distribution as the joint distribution of the n random variables X_s. Therefore, by subadditivity of the entropy function, we have

$$H(X) \le \sum_{s\in S} H(X_s) = \sum_{s\in S} h(p_s).$$

Hence, we get the upper bound

$$r_S(x) \le 2^h,$$

where h is the maximum value of $\sum_{s\in S} h(p_s)$ over all choices of $(p_s)_{s\in S}$ satisfying $\sum_{s\in S} sp_s \le x$. In particular, for $S = \{1/m : m \in [n]\}$, we obtain that the number of subsets $A \subseteq [n]$ with $s(A) \le x$ is at most $2^{H(P(x))}$, as claimed.

We now turn to the lower bound. Consider independent Bernoulli random variables Y_m for $m \in [n]$ satisfying $Y_m = 1$ with probability p_m and $Y_m = 0$ otherwise. Let $Y = (Y_m)_{m\in[n]}$, $Z = \sum_{m\in[n]} Y_m/m$ and E be the indicator of the event $Z \le x$. Recall that, for $a \in \{0,1\}$, the conditional entropy $H(Y|E=a) = -\sum_{y\in\{0,1\}^n} \Pr[Y = y|E = a] \log \Pr[Y = y|E = a]$ and $H(Y|E) = \sum_{a\in\{0,1\}} \Pr[E = a]H(Y|E = a)$. Since E is determined by Y, we have

$$H(Y) = H(Y, E) = H(Y|E) + H(E) = H(E) + \sum_{a\in\{0,1\}} \Pr[E = a]H(Y|E = a).$$

We thus have

$$H(Y|E = 1) = \frac{1}{\Pr[Z \le x]} \left(H(Y) - H(E) - \Pr[Z > x]H(Y|E = 0)\right). \tag{2}$$

Let $c = c_{x,n}$ as in Lemma 3. By that lemma, the random variable Z has variance

$$\sum_{m=1}^{n} \left(\frac{1}{(1 + e^{cn/m})m^2} - \frac{1}{(1 + e^{cn/m})^2 m^2} \right) = \Theta\left(\sum_{m=1}^{n} \frac{e^{-cn/m}}{m^2} \right) = \Theta\left(\frac{1}{cn} \right). \tag{3}$$

To see the last bound, observe that $\sum_{m=cn/(k+1)}^{cn/k} \frac{e^{-cn/m}}{m^2} = \Theta(\frac{e^{-k}}{cn})$. Similarly, for $1 \le k \le 1/c$, $\sum_{m=cnk}^{cn(k+1)} \frac{e^{-cn/m}}{m^2} = \Theta(\frac{e^{-1/k}}{cnk^2})$. The desired bound follows from summing these estimates over k.

By a similar argument, the sum of the centered third moments of the Y_i is

$$\begin{aligned} \sum_{m=1}^{n} \left[\frac{1}{1 + e^{cn/m}} \left(\frac{1}{m} - \frac{1}{m(1 + e^{cn/m})} \right)^3 + \left(1 - \frac{1}{1 + e^{cn/m}} \right) \left(-\frac{1}{m(1 + e^{cn/m})} \right)^3 \right] &= O\left(\sum_{m=1}^{n} \frac{e^{-cn/m}}{m^3} \right) \\ &= O\left(\frac{1}{(cn)^2} \right). \end{aligned} \tag{4}$$

The Berry–Esseen bound, Lemma 4, then yields that, for $g \sim \mathcal{N}(0, 1)$,

$$\Pr(E = 1) = \Pr[Z \le x] = \Pr[g \le 0] + O((cn)^{-1/2}) = 1/2 + O((cn)^{-1/2}), \tag{5}$$

where we used that $\mathbb{E}[Z] = x$, together with (3) and (4).

We next bound $H(Y_1, \ldots, Y_n | E = 0) \le \sum_{m=1}^{n} H(Y_m | E = 0)$. To bound the summands, we note by Bayes' rule that

$$\Pr(Y_m = 1 | E = 0) = \frac{\Pr(E = 0 | Y_m = 1) \Pr(Y_m = 1)}{\Pr(E = 0)}$$

and we will use a similar argument with the Berry–Esseen bound to show that $\Pr(Y_m = 1 | E = 0)$ is close to $\Pr(Y_m = 1)$. Indeed, the calculations above similarly yield that the random variable $Z'_m = Z - Y_m + 1/m$ is a sum of independent random variables with $\mathbb{E}Z'_m = x + \frac{1}{m} - \frac{1}{m(1+e^{cn/m})}$, $\mathrm{Var}(Z'_m) = \Theta(1/(cn))$ and the sum of centered third moments $O(1/(cn)^2)$. By Lemma 4, for $g \sim \mathcal{N}(0,1)$,

$$\begin{aligned}
\Pr(E = 0 | Y_m = 1) &= \Pr(Z'_m > x) \\
&= \Pr\left(g > -\frac{1/m - 1/(m(1+e^{cn/m}))}{\mathrm{Var}(Z'_m)^{1/2}}\right) + O((cn)^{-1/2}) \\
&= 1/2 + O\left((cn)^{-1/2} + \frac{\sqrt{cn}}{m}\right),
\end{aligned}$$

assuming that $m > 10\sqrt{cn}$ for the last bound, where we used the simple estimate $\Pr(g > z) = \frac{1}{2} + O(z)$ for $|z| \le 1$. Therefore,

$$\left|\frac{\Pr(E = 0 | Y_m = 1)}{\Pr(E = 0)} - 1\right| = \left|\frac{1/2 + O\left((cn)^{-1/2} + \frac{\sqrt{cn}}{m}\right)}{1/2 + O((cn)^{-1/2})} - 1\right| \le O\left(\frac{\sqrt{cn}}{m} + \frac{1}{\sqrt{cn}}\right).$$

Thus, by Bayes' rule,

$$\frac{|\Pr(Y_m = 1 | E = 0) - \Pr(Y_m = 1)|}{\Pr(Y_m = 1)} \le O\left(\frac{\sqrt{cn}}{m} + \frac{1}{\sqrt{cn}}\right). \tag{6}$$

From (6), we have

$$\Pr(Y_m = 1)\left(1 - O\left(\frac{\sqrt{cn}}{m} + \frac{1}{\sqrt{cn}}\right)\right) \le \Pr(Y_m = 1 | E = 0) \le \Pr(Y_m = 1)\left(1 + O\left(\frac{\sqrt{cn}}{m} + \frac{1}{\sqrt{cn}}\right)\right).$$

Since $h'(p) = \log \frac{1-p}{p} \le \log \frac{1}{p}$, we have that

$$\begin{aligned}
&h(\Pr(Y_m = 1 | E = 0)) \\
&\le h(\Pr(Y_m = 1)) + \left(\log \frac{1}{\Pr(Y_m = 1)\left(1 - O\left(\frac{\sqrt{cn}}{m} + \frac{1}{\sqrt{cn}}\right)\right)}\right) |\Pr(Y_m = 1 | E = 0) - \Pr(Y_m = 1)| \\
&\le h(\Pr(Y_m = 1)) + \left(\log \frac{1}{\Pr(Y_m = 1)\left(1 - O\left(\frac{\sqrt{cn}}{m} + \frac{1}{\sqrt{cn}}\right)\right)}\right) \cdot O\left(\frac{\sqrt{cn}}{m} + \frac{1}{\sqrt{cn}}\right) \Pr(Y_m = 1) \\
&\le h(\Pr(Y_m = 1)) + O\left(\frac{\sqrt{cn}}{m} + \frac{1}{\sqrt{cn}}\right) \Pr(Y_m = 1) \log \frac{1}{\Pr(Y_m = 1)},
\end{aligned}$$

where in the last inequality we used that $\Pr(Y_m = 1) \le 1/2$, so $\log \frac{1}{\Pr(Y_m=1)\left(1-O\left(\frac{\sqrt{cn}}{m}+\frac{1}{\sqrt{cn}}\right)\right)} = O\left(\log \frac{1}{\Pr(Y_m=1)}\right)$. Therefore,

$$\begin{aligned}
H(Y_1, \ldots, Y_n | E = 0) &\le \sum_{m=1}^{n} H(Y_m | E = 0) \\
&\le 10\sqrt{cn} + \sum_{m > 10\sqrt{cn}} h(\Pr(Y_m = 1 | E = 0)) \\
&\le 10\sqrt{cn} + \sum_{m > 10\sqrt{cn}} \left(H(Y_m) + O\left(\frac{\sqrt{cn}}{m} + \frac{1}{\sqrt{cn}}\right) \Pr(Y_m = 1) \log \frac{1}{\Pr(Y_m = 1)}\right) \\
&\le H(Y_1, \ldots, Y_n) + 10\sqrt{cn} + \sum_{m > 10\sqrt{cn}} O\left(\frac{\sqrt{cn}}{m} + \frac{1}{\sqrt{cn}}\right) \Pr(Y_m = 1) \log \frac{1}{\Pr(Y_m = 1)}
\end{aligned}$$

$$\leq H(Y_1,\dots,Y_n) + 10\sqrt{cn} + \sum_{m>10\sqrt{cn}} O\left(\frac{\sqrt{cn}}{m} + \frac{1}{\sqrt{cn}}\right)\frac{cn/m}{1+e^{cn/m}}$$
$$\leq H(Y_1,\dots,Y_n) + O(\sqrt{n/c}).$$

Again, for the last estimate, we note that $\sum_{m=cn/(k+1)}^{cn/k}\left(\frac{\sqrt{cn}}{m}+\frac{1}{\sqrt{cn}}\right)\frac{cn/m}{1+e^{cn/m}} \leq O\left(\frac{cn}{k^2}e^{-k}\frac{k^2}{\sqrt{cn}}\right) = O(e^{-k}\sqrt{cn})$ and, for $1 \leq k \leq 1/c$, $\sum_{m=cnk}^{cn(k+1)}\left(\frac{\sqrt{cn}}{m}+\frac{1}{\sqrt{cn}}\right)\frac{cn/m}{1+e^{cn/m}} \leq O\left(cn\frac{1}{k\sqrt{cn}}\right) \leq O\left(\frac{1}{k}\sqrt{cn}\right)$. Summing over k, we thus have

$$\sum_{m>10\sqrt{cn}} O\left(\frac{\sqrt{cn}}{m} + \frac{1}{\sqrt{cn}}\right)\frac{cn/m}{1+e^{cn/m}} = O(\sqrt{n/c}).$$

Combining with (2), and noting that $H(E) \leq 1$ and $\Pr(Z \leq x) = 1/2 + O((cn)^{-1/2})$ by (5), we obtain that

$$H(Y_1,\dots,Y_n|E=1) \geq (1-O((cn)^{-1/2}))(H(P(x)) - O(\sqrt{n/c})) = H(P(x)) - O(\sqrt{n/c}).$$

Using that for any random variable X we have $H(X) \leq \log|\mathrm{supp}(X)|$, we obtain that the number of subsets $A \subseteq [n]$ with $s(A) \leq x$ is at least $2^{H(P(x))-O(\sqrt{n/c})}$. This implies that the number of subsets $A \subseteq [n]\setminus U$ with $s(A) \leq x$ is at least

$$2^{-(n-|U|)}2^{H(P(x))-O(\sqrt{n/c})},$$

as required. □

3 Subset sums of modular inverses

The main technical tool we still need is the following result, which says that if q is a large prime power and we take a dense subset I of the interval $[q^\varepsilon, 2q^\varepsilon]$, then every residue class mod q can be written as the sum of a small number of reciprocals of elements of I. Roughly speaking, this allows us to cancel out any particular prime power from the denominator of a fraction in the absorption step of the proof of Theorem 1. Given a set A of integers, we will use the notation $\Sigma^{[s]}(A)$ for the collection of sums of subsets of A of size at most s.

Theorem 5. *Let $\delta, \varepsilon > 0$ and let q be a prime power which is sufficiently large in terms of δ, ε. If I is a subset of $[q^\varepsilon, 2q^\varepsilon]$ consisting of elements coprime to q with $|I| \geq \delta q^\varepsilon$, then $\Sigma^{[s]}(I^{-1}) \pmod q = \mathbb{Z}_q$ for $s = q^{\varepsilon/2}$.*

In the proof of Theorem 5, we will make use of the following key result from [5]. Recall that a *generalized arithmetic progression (henceforth GAP) P of dimension k* is a set of integers $\{x_0 + \ell_1 x_1 + \ell_2 x_2 + \cdots + \ell_k x_k | 0 \leq \ell_1 < L_1, \dots, 0 \leq \ell_d < L_k\}$. A GAP is called *proper* if it has size exactly $L_1 L_2 \cdots L_k$. We say that P is *homogeneous* if x_0 divides $x_1, \dots, x_k$. For a natural number t, we define tP to be the t-fold sumset of P, while if t is a positive real number which is not an integer and $P = \{\sum_{i=1}^k n_i x_i : a_i \leq n_i \leq b_i\}$ is a homogeneous GAP, we can generalize the definition by setting $tP = \{\sum_{i=1}^k n_i x_i : ta_i \leq n_i \leq tb_i\}$.

Theorem 6. *For any $\beta > 1$ and $0 < \eta < 1$, there are positive constants c and k such that the following holds. Let A be a subset of $[n]$ of size m with $n \leq m^\beta$ and let $s \in [m^\eta, cm/\log m]$. Then there exists a subset $\hat{A}$ of A of size at least $m - c^{-1}s\log m$ and a proper GAP P of dimension at most k such that $\hat{A} \cup \{0\}$ is a subset of P. Furthermore, there exists $A' \subseteq \hat{A}$ of size at most s such that $\Sigma(A')$ contains a homogeneous translate of csP, where csP is proper.*

We will also need the following simple variant of Dirichlet's simultaneous approximation theorem. For a residue class $i \pmod q$, we use the notation $\overline{i}$ for the unique integer in $(-q/2, q/2]$ congruent to i modulo q.

Lemma 7. *Given a prime power q, integers $d_1, \dots, d_k$ coprime to q and positive integers $a_1, \dots, a_k$ such that $\prod_{i=1}^k a_i = A$, there exists a positive integer $T < q$ and integers $d_1', \dots, d_k'$ such that $Td_i = d_i' \pmod q$ and $|d_i'| \leq 2(q/a_i)\cdot(A/q)^{1/k}$ for all $i \in [k]$.*

Proof. Let $b_i = 2(q/a_i)\cdot(A/q)^{1/k}$. Note that $\lfloor \overline{sd_i}/b_i \rfloor$ takes at most q/b_i values as s ranges over $\mathbb{Z}_q$. By the pigeonhole principle, there exist distinct $s \neq s'$ in $\mathbb{Z}_q$ such that $\lfloor \overline{sd_i}/b_i \rfloor = \lfloor \overline{s'd_i}/b_i \rfloor$ for all $i \in [k]$, since $\prod_{i=1}^k \frac{q}{b_i} = (q/A)2^{-k}\prod_{i=1}^k a_i < q$. Letting $T = s' - s$, we then have that $|\overline{Td_i}| \leq b_i = 2(q/a_i)\cdot(A/q)^{1/k}$. □

We now proceed to the proof of Theorem 5. The basic idea is to use Theorem 6 to argue that there is a large subset J of the set of inverses I^{-1} which is contained in a proper GAP P of bounded dimension k such that $\Sigma^{[s]}(J)$ contains a proper translate of csP. We then exploit the nature of the set of inverses to argue that k must in fact be 1, that is, P is simply a progression, from which the required result quickly follows.

Proof of Theorem 5. Let $s = q^{\varepsilon/2}$. Let $\overline{I}$ and $\overline{I^{-1}}$ denote the set of integer representations (in $(-q/2, q/2]$) of I and I^{-1}. By Theorem 6, there is c depending only on ε such that we can find $J \subseteq \overline{I^{-1}}$ of size at least $|I| - c^{-1}s\log|I| = (1-o(1))|I|$ and a proper GAP P of dimension $k = O_\varepsilon(1)$ such that $J \cup \{0\} \subseteq P$ and $\Sigma^{[s]}(J)$ contains a translate of csP which is proper.

By expanding P by a factor of up to 2^k if necessary, we can write $P = \sum_{u=1}^k [-a_u, a_u]d_u$. Let $A = \prod_{u=1}^k a_u$. With these a_u and d_u, we apply Lemma 7 to find a value of T satisfying the conclusions of that lemma and let $T \cdot P = \{tx : x \in P\}$. Note that, for any $j \in T \cdot P$, $|\overline{j}| \le \sum_{u=1}^k a_u |d_u'| \le 2kq(A/q)^{1/k}$.

Claim. *Let N denote the number of solutions to the equation $\overline{i} \cdot \overline{j} = T \pmod q$ with $i \in \overline{I}$ and $j \in T \cdot J$. Then*

$$|I|/2 \le N < q^{C/\log\log q} \cdot 8kq^{\varepsilon}(A/q)^{1/k}. \tag{7}$$

We first complete the proof of Theorem 5 assuming the claim. From (7) and the assumption that $|I| \ge \delta q^\varepsilon$, we deduce that

$$A \ge \delta^k (16k)^{-k} q^{1 - Ck/\log\log q}.$$

On the other hand, since $\Sigma^{[s]}(J)$ contains a translate of csP with csP proper and $\Sigma^{[s]}(J) \subseteq (-sq/2, sq]$, we have that

$$c^k s^k A \le c^k s^k |P| = |csP| \le |\Sigma^{[s]}(J)| \le sq.$$

Hence, $A \le c^{-k} q s^{1-k} = c^{-k} q^{1-(k-1)\varepsilon/2}$. Provided q is sufficiently large in terms of δ, ε, these two estimates on A together imply that $k = 1$. Therefore, csP is an arithmetic progression of length $\Omega(As) > q$. Furthermore, P must have common difference coprime with q as $J \subseteq \overline{I^{-1}}$ is contained in P. Hence, any translate of csP covers all residue classes in $\mathbb{Z}_q$. This finishes the proof of the theorem assuming the claim.

It remains to verify the claim. Since each $j_0 \in J$ has $j_0^{-1} \in I \pmod q$, the number of solutions to $\overline{i} \cdot \overline{j} = T \pmod q$ with $i \in \overline{I}$, $j = T \cdot j_0 \in T \cdot J \subseteq T \cdot P$ is at least $|J| = (1-o(1))|I| \ge |I|/2$. Since $I \subseteq [q^\varepsilon, 2q^\varepsilon]$, we have that

$$|\overline{i} \cdot \overline{j}| \le 2q^\varepsilon \cdot 2kq(A/q)^{1/k}.$$

As such, if $\overline{i} \cdot \overline{j} = T \pmod q$, then $\overline{i} \cdot \overline{j} = qx + T$, where $0 \le |x| \le 4kq^\varepsilon(A/q)^{1/k}$. But the number of solutions to the equation $\overline{i} \cdot \overline{j} = qx + T$ with $0 \le |x| \le 4kq^\varepsilon(A/q)^{1/k}$ is bounded above by

$$\sum_{|x| \le 4kq^\varepsilon(A/q)^{1/k}} \tau(qx+T) < q^{C/\log\log q} \cdot 8kq^\varepsilon(A/q)^{1/k},$$

where $\tau(n)$ denotes the number of divisors of n and we have used the standard bound $\tau(n) \le q^{C/\log\log q}$ for an absolute constant C and all $n \le q^2$. This completes the proof of the claim. □

4 Absorption

We are now ready to prove Theorem 1 in the following explicit form. We recall that a positive integer n is *t-smooth* if all of its prime factors are at most t and *t-powersmooth* if all of its prime power factors are at most t.

Theorem 8. *Let $\varepsilon > 0$ be sufficiently small. Then there exists $\xi > 0$ such that if $x \le \xi \ln n$ is a rational whose denominator is $(n^{1-\varepsilon}/2)$-powersmooth, then the number of subsets $A \subseteq [n]$ with $x = \sum_{a \in A} 1/a$ is at least $2^{c_x n - c_\varepsilon n}$, where $c_\varepsilon \to 0$ as $\varepsilon \to 0$ and*

$$c_x := \int_0^1 h\left(\frac{1}{1+e^{\lambda/y}}\right) dy$$

with λ the unique real number such that

$$\int_0^1 \frac{1}{y(1+e^{\lambda/y})} dy = x.$$

We first record a simple lemma guaranteeing that most integers at most n are $(n^{1-\varepsilon}/2)$-powersmooth.

Lemma 9. *For δ sufficiently small, n sufficiently large in terms of δ and $t = n^{1-\delta}$, at least $(1-2\delta)n$ positive integers at most n are t-powersmooth.*

Proof. It is well known that if $t = n^u$, the number of t-smooth numbers up to n is asymptotic to $\psi(u)n$, where ψ is the Dickman function taking values in $(0,1)$ for $u \in (0,1)$. Moreover, for $u > 1/2$, $\psi(u) = 1 + \ln u$. Thus, for $u = 1-\delta$ and $t = n^{1-\delta}$, at least $(1 + \ln(1-\delta) - o(1))n \geq (1 - \frac{3}{2}\delta)n$ positive integers at most n are $n^{1-\delta}$-smooth.

Among the t-smooth numbers, the only ones that are not t-powersmooth are those divisible by a prime power p^α where $p \leq t$ but $p^\alpha > t$. Using the prime number theorem, we may upper bound the total count of such exceptional smooth numbers by

$$\sum_{p \leq t} \left\lfloor \frac{n}{t} \right\rfloor = \pi(t) \left\lfloor \frac{n}{t} \right\rfloor = o(n).$$

Hence, at least $(1 - \frac{3}{2}\delta)n - o(n) \geq (1-2\delta)n$ positive integers at most n are t-powersmooth, as required. □

Finally, we prove Theorem 8. Recall the notation that, for $A \subseteq [n]$, $s(A) = \sum_{a \in A} 1/a$.

Proof of Theorem 8. By choosing c_ε suitably, we can assume that n is sufficiently large in terms of ε. Let L be sufficiently large, assuming in particular that Theorem 5 applies for ε as in the statement of the theorem, $\delta = \frac{1}{2}$ and all $q > L$. Let K denote the least common multiple of all prime powers at most L. We first reserve the set $\mathcal{R}$ of multiples of K in $[n]$. Let $\mathcal{P}(q) = q \cdot [q^\varepsilon, 2q^\varepsilon] \setminus \mathcal{R}$ and $\mathcal{P} = \bigcup_q \mathcal{P}(q)$, where q ranges over all prime powers at most $n^{1-\varepsilon}/2$. Here $q \cdot S = \{qs : s \in S\}$ and the notation $[q^\varepsilon, 2q^\varepsilon]$ refers to the set of integers in this interval. Let $\mathcal{S}$ denote the set of $(n^{1-\varepsilon}/2)$-powersmooth numbers at most n and $\mathcal{U} = \mathcal{S} \setminus (\mathcal{R} \cup \mathcal{P})$. Lemma 9 implies that $|\mathcal{S}| \geq (1 - O(\varepsilon))n$ and we also have that $|\mathcal{P} \cap [n]| \leq \sum_{q \leq n^{1-\varepsilon}} 2q^\varepsilon \leq 4\frac{n}{\log n}$. Thus,

$$n - |\mathcal{U}| \leq n/K + O(\varepsilon n). \tag{8}$$

Let $\eta > 0$ be a constant to be chosen later. By Lemma 2, applied with x replaced by $(1-\eta)x$, we can find many subsets of $\mathcal{U}$ whose sums of reciprocals are at most $(1-\eta)x$. Indeed, the number of such subsets is at least

$$2^{H(P((1-\eta)x)) - (n - |\mathcal{U}|) - O(\sqrt{n/c_{(1-\eta)x,n}})}.$$

Fix one such sum corresponding to a set $A_0 \subseteq [n]$, and let $x_0 = x - s(A_0) \geq \eta x$. Consider the following procedure, where at each step i we have a real number x_i and a set A_i for which $s(A_i) = x_i$:

1. In decreasing order over the prime powers larger than L, consider the largest prime power $q = q_i \leq n^{1-\varepsilon}/2$ of a prime $p = p_i$ which appears as a factor of the denominator of x_i. We then find $B_i \subseteq (1/q) \cdot \mathcal{P}(q)$ of size at most $q^{\varepsilon/2}$ such that, for $x_i = \frac{u_i}{v_i}$ with u_i, v_i coprime, $s(B_i) = -\frac{u_i}{v_i/q} \pmod q$. We say that step i succeeds if we can find such a B_i. If it does succeed, we update $x_{i+1} = x_i - s(q \cdot B_i)$ and $A_{i+1} = A_i \cup q \cdot B_i$, noting by our choice that no nonzero power of p divides the denominator of x_{i+1}. Furthermore, any new prime power divisor of x_{i+1} is at most $2q^\varepsilon$.

2. We iterate until all the prime powers $q_i > L$ have been processed. At this point, the final output x_f is a rational number whose denominator is L-powersmooth. We then find a subset of the reservoir $\mathcal{R}$ whose sum of inverses is equal to x_f.

The following claim guarantees that the procedure above succeeds.

Claim. *For each i, step i succeeds. Furthermore, $s(q_i \cdot B_i) \leq q_i^{-1-\varepsilon/2}$ and, for some absolute constant $C > 0$,*

$$|x_f - x_0| \leq C\varepsilon^{-1} L^{-\varepsilon/2}. \tag{9}$$

Proof. Theorem 5 implies immediately that step i always succeeds. Furthermore, by our choice of B_i,

$$s(q_i \cdot B_i) \leq \frac{q_i^{\varepsilon/2}}{q_i^{1+\varepsilon}} = q_i^{-1-\varepsilon/2}.$$

The estimate (9) follows since the q_i are distinct integers between L and n, so

$$|x_f - x_0| = \sum_i s(q_i \cdot B_i) \leq \sum_{L < q_i < n} q_i^{-1-\varepsilon/2} \leq \sum_{L < m < n} m^{-1-\varepsilon/2} < O(\varepsilon^{-1} L^{-\varepsilon/2}),$$

as required. □

We ensure that $L, \eta > 0$ are chosen (depending on ε) so that $C\varepsilon^{-1}L^{-\varepsilon/2} < \eta x/2$. From (9), we have that $x \geq x_f$ and $x - x_f$ is a positive rational number at most x whose denominator is L-powersmooth. As such, we have that $K(x - x_f)$ is a positive integer with $K(x - x_f) \leq Kx$. We now note that there exists a subset $D \subseteq [n/K]$ such that $s(D) = K(x - x_f)$, where we use the assumption that $x \leq \xi \ln n$ for ξ chosen sufficiently small in ε, so that $Kx < \varepsilon \log(n/K)$. To see that this is the case, one may, for example, make use of Croot's result [6] that one can always be written as the sum of reciprocals of numbers from any interval of the form $[t, (e + o(1))t]$. This allows us to iteratively remove $K(x - x_f)$ disjoint subsets from $[n/K]$, the sum of the reciprocals of each of which is one. We then set their union to be D, noting that $K \cdot D \subseteq \mathcal{R}$ is disjoint from $\mathcal{U}$ and $\mathcal{P}$.

We then have $x = \sum_{d \in D} \frac{1}{Kd} + \sum_i s(q_i \cdot B_i) + s(A_0)$ and the number of such distinct representations is at least the number of choices for A_0, which is bounded below by

$$2^{H(P((1-\eta)x)) - (n - |\mathcal{U}|) - O(\sqrt{n/c_{(1-\eta)x,n}})} \geq 2^{H(P(x)) - c_\varepsilon n},$$

for an appropriate constant c_ε with $c_\varepsilon \to 0$ as $\varepsilon \to 0$, where we have used Lemma 2 and (8). This completes the proof of Theorem 8. □

Note added. As we completed this paper, we learned that a result similar to our Theorem 1 was obtained simultaneously and independently, though using rather different methods, by Yang P. Liu and Mehtaab Sawhney [9].

Acknowledgements. We are grateful to the American Institute of Mathematics for hosting the SQuaREs project at which this work was initiated. Research supported by NSF Awards DMS-2054452 and DMS-2348859 (David Conlon), NSF Award DMS-2154129 (Jacob Fox), NSF Award DMS-2103154 (Xiaoyu He), NSF Awards DMS-1952767 and DMS-2153576 (Dhruv Mubayi), a Clay Research Fellowship and a Stanford Science Fellowship (Huy Tuan Pham), an NSF CAREER Award and NSF Awards DMS-1952786 and DMS-2246847 (Andrew Suk) and NSF Award DMS-1800332 (Jacques Verstraëte).

References

[1] A. C. Berry, The accuracy of the Gaussian approximation to the sum of independent variates, *Trans. Amer. Math. Soc.* **49** (1941), 122–136. 2

[2] T. F. Bloom, On a density conjecture about unit fractions, preprint available at arXiv:2112.03726 [math.NT]. 1

[3] T. F. Bloom, www.erdosproblems.com, March 2024. 1

[4] T. F. Bloom and C. Elsholtz, Egyptian fractions, *Nieuw Arch. Wiskd.* **23** (2022), 237–245. 1

[5] D. Conlon, J. Fox and H. T. Pham, Homogeneous structures in subset sums and non-averaging sets, preprint available at arXiv:2311.01416 [math.CO]. 1, 3

[6] E. S. Croot, On unit fractions with denominators in short intervals, *Acta Arith.* **99** (2001), 99–114. 4

[7] C.-G. Esseen, On the Liapunoff limit of error in the theory of probability, *Arkiv för Matematik, Astronomi och Fysik* **A28** (1942), 1–19. 2

[8] P. Erdős and R. L. Graham, Old and new problems and results in combinatorial number theory, Monogr. Enseign. Math., 28, Université de Genève, L'Enseignement Mathématique, Geneva, 1980, 128 pp. 1

[9] Y. P. Liu and M. Sawhney, On further questions regarding unit fractions, preprint available at arXiv:2404.07113 [math.NT]. 4

[10] I. G. Shevtsova, An improvement of convergence rate estimates in the Lyapunov theorem, *Dokl. Math.* **82** (2010), 862–864. 2

[11] S. Steinerberger, On a problem involving unit fractions, preprint available at arXiv:2403.17041 [math.CO]. 1