Bit-Probe Lower Bounds for Succinct Data Structures

Emanuele Viola[*]
Northeastern University

December 3, 2008

Abstract

We prove lower bounds on the redundancy necessary to represent a set S of objects using a number of bits close to the information-theoretic minimum $\log_2 |S|$, while answering various queries by probing few bits. Our main results are:

- To represent n ternary values $t \in \{0,1,2\}^n$ in terms of u bits $b \in \{0,1\}^u$ while accessing a single value $t_i \in \{0,1,2\}$ by probing q bits of b, one needs

  $$u \geq (\log_2 3)n + n/2^{O(q)}.$$

  This matches an exciting representation by Pătraşcu (FOCS 2008) where $u \leq (\log_2 3)n + n/2^{q^{\Omega(1)}}$. We also note that results on logarithmic forms imply the lower bound $u \geq (\log_2 3)n + n/\log^{O(1)} n$ if we access t_i by probing one cell of $\log n$ bits.

- To represent sets of size $n/3$ from a universe of n elements in terms of u bits $b \in \{0,1\}^u$ while answering membership queries by probing q bits of b, one needs

  $$u \geq \log_2 \binom{n}{n/3} + n/2^{O(q)} - \log n.$$

Both results above hold even if the probe locations are determined adaptively.

Ours are the first lower bounds for these fundamental problems; we obtain them drawing on ideas used in lower bounds for locally decodable codes.

[*] viola@ccs.neu.edu

1 Introduction

A *succinct data structure* is an encoding $Enc : S \to \{0,1\}^u$ of a set S of objects that allows for efficient answers to various queries, while at the same time using space close to the information-theoretic minimum: $u = \log_2 |S| + r$ for a small *redundancy* $r \ll \log_2 |S|$. There has been considerable interest and progress in exhibiting such data structures, see for example [Mit, Pag, GRRR, Păt] and the references therein. Less progress seems to have been made on negative results, a.k.a. lower bounds, with the few known results applying either under artificial restrictions [GGG+, GRR] or to somewhat contrived problems [GM].

In this work, we prove new lower bounds for fundamental problems of succinct data structures, in some cases matching the known upper bounds. We now discuss a couple of problems and present our main results.

1.1 Representing ternary values using bits

Following Pătraşcu [Păt], consider the problem of representing an array of n ternary values $t = (t_1, \ldots, t_n) \in \{0,1,2\}^n$ in terms of u bits $b \in \{0,1\}^u$. This is a fundamental problem, since data is often arranged in tuples of elements from a domain that is not a power of 2 (e.g., students' grades are in $\{A, B, C, D, F\}$). For representing $t \in \{0,1,2\}^n$ in $\{0,1\}^u$, the information-theoretic minimum is $u = \lceil (\log_2 3)n \rceil$. We can match this minimum using arithmetic coding: view t as an integer between 0 and $3^n - 1$, and write down its u-bit binary representation. The drawback is that to access a value $t_i \in \{0,1,2\}$ we have to read the whole representation, i.e., probe u bits. At the other end of the spectrum, we can represent each ternary value $t_i \in \{0,1,2\}$ using 2 bits. Here we access each t_i by probing just 2 bits, but use $u = 2n \gg (\log_2 3)n$ bits of space.

A tradeoff between these two extremes is obtained by using arithmetic coding for each block of k ternary values. Now, to access a value $t_i \in \{0,1,2\}$, we probe the $q := \lceil (\log_2 3)k \rceil$ bits of the encoding of the block containing it, and the space used is

$$u = \lceil (\log_2 3)k \rceil \cdot n/k = (\log_2 3)n + n/k^c, \tag{1}$$

where $c \geq 1$ is immediate, and results on logarithmic forms discussed in §4 imply that c is bounded from above by an absolute constant. So this approach gives a *polynomial* tradeoff between the number $q = \Theta(k)$ of probes and the redundancy $n/k^c = n/q^{\Theta(1)}$ of the structure.

A recent, exciting work by Pătraşcu [Păt, Th. 4 for $w := 2t$] gives a better, *exponential* tradeoff between the number q of bits probed and the redundancy:

$$u \leq (\log_2 3)n + n/2^{q^{\Omega(1)}}. \tag{2}$$

In this work we prove the first lower bound for this problem, establishing that Pătraşcu's exponential tradeoff (2) is optimal up to the constant in the "$\Omega(1)$."

Theorem 1.1 (Lower bound for representing ternary values). *To represent $\{0,1,2\}^n$ in $\{0,1\}^u$ supporting single-element access by probing q bits, one needs*

$$u \geq (\log_2 3)n + n/2^{6q+22}.$$

1.2 Representing sets using bits

The *dictionary problem* is another basic problem in data structures which asks to represent a set of size ℓ from a universe of n elements in terms of u bits $b \in \{0,1\}^u$ so that membership queries can be answered efficiently. The classic work by Minsky and Papert [MP] already studies representation of sets where membership can be determined by probing a few bits (on average). More recently, Buhrman, Miltersen, Radhakrishnan, and Venkatesh [BMRV] give a surprising representation whose space is within a constant factor of the information-theoretic minimum $\log_2 \binom{n}{\ell}$, and membership is determined by reading just one bit (with high probability). In terms of lower bounds, they prove that, to represent sets of size ℓ from a universe of n elements in terms of u bits, answering membership queries by probing q bits, one needs

$$\binom{n}{\ell} \le 2^{\ell \cdot q} \cdot \binom{u}{\ell \cdot q}.$$

Their lower bound is interesting when $\ell \le n^{1-\Omega(1)}$, but gives little information when $\ell = \theta(n)$. For example, it gives nothing for $\ell = n/3$ and $q = 3$. In fact, no general lower bound seems to have been known for this "close to capacity" regime $\ell = \theta(n)$.

With a proof that is very similar to that of Theorem 1.1, in this work we prove the following lower bound.

Theorem 1.2 (Lower bound for representing sets)**.** *For all sufficiently large n divisible by 3, to represent $S := \{x : x \in \{0,1\}^n, \sum_i x_i = n/3\}$ in $\{0,1\}^u$ answering membership queries by probing q bits, one needs*

$$u \ge \log_2 |S| + n/2^{6q+22} - \log_2 n.$$

To mention other upper bounds for the dictionary problem, we need to distinguish between two popular computational models. The model discussed until now is usually called *bit-probe*, because each probe in the data structure returns a bit. A more general model is the so-called *cell-probe* model, where the memory is divided in cells of $\Theta(\log n)$ bits, and each probe returns the content of an entire cell (see Miltersen's survey [Mil] for background). In the cell-probe model, there are efficient, succinct data structures for the dictionary problem: building on the results by Pagh [Pag], Pătraşcu [Păt] gives a representation of sets of size ℓ from a universe of n elements that uses space

$$u \le \log_2 \binom{n}{\ell} + n/\log^c n, \tag{3}$$

and where membership queries are answered by probing q cells of $\Theta(\log n)$ bits each, where $q = q(c)$ depends only on c. While our main results only apply to the bit-probe model, they also give some indication that beating the construction in (3) may be difficult: [Păt] obtains (3) and the succinct representation of ternary values (2) using similar techniques, but in light of our Theorem 1.1, to beat (3) one should use ideas that do not apply to (2). In §4 we discuss the cell-probe model more.

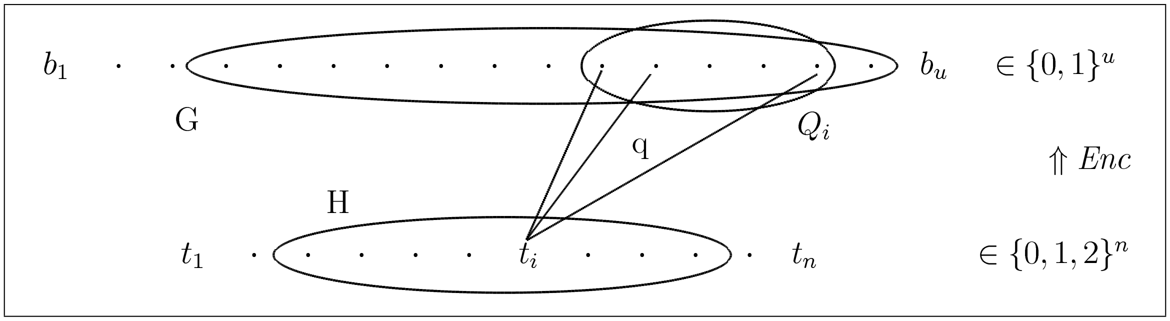

Figure 1: Lower bound for representing ternary values $\{0,1,2\}^n$ in $\{0,1\}^u$.

1.3 Techniques

In this section we discuss the techniques we use to prove our lower bounds. We focus on the problem of representing ternary values $t \in \{0,1,2\}^n$ in $\{0,1\}^u$ (Theorem 1.1) because it is clean; later we discuss how to obtain the lower bound for dictionaries as well. We first explain the proof under the assumption that the probe locations are non-adaptive, i.e., only depend on the ternary value to be accessed but not on the results of previous probes. Later we address adaptivity.

Intuitively, representing ternary values using bits is difficult because ($\star$) *ternary values are not binary, in the sense that they have a number of combinations which is not a power of* 2. Our proof formalizes precisely this intuition. We now explain our proof, and we refer the reader to Figure 1 for an illustration of our reasoning. Suppose that we represent $\{0,1,2\}^n$ in $\{0,1\}^u$ were $u = (\log_2 3)n + r$ is very close to the information-theoretic minimum: $r \ll n$. Let us choose $t \in \{0,1,2\}^n$ uniformly at random, and consider the encoding $Enc(t) \in \{0,1\}^u$ of t. The encoding is obviously one-to-one (since every ternary value can be recovered), thus $Enc(t)$ is uniformly distributed over 3^n elements of $\{0,1\}^u$. Since $2^u \approx 3^n$, the entropy of $Enc(t)$ is very close to the maximum u. Therefore we can apply a relatively standard information-theoretic Lemma 2.2 which states that the random variable $b = Enc(t) \in \{0,1\}^u$ is approximately uniform, in the sense that there is a large set of indices $G \subseteq [u]$ such that for any q indices $\{i_1,\dots,i_q\} \subseteq G$, the distribution of $(b_{i_1},\dots,b_{i_q})$ is essentially uniform over $\{0,1\}^q$ [Raz, EIRS, SV].

Suppose now that we can decode a ternary value t_i as a function d_i of q probes $Q_i := \{i_1,\dots,i_q\}$:

$$t_i = d_i(b_{i_1},\dots,b_{i_q}). \tag{4}$$

If the probes $Q_i = \{i_1,\dots,i_q\}$ are all in G, then $d_i(b_{i_1},\dots,b_{i_q})$ is essentially distributed like $d(U)$ where U is uniform over $\{0,1\}^q$. Now we can return to the intuition at the beginning of this section ($\star$): A uniformly distributed ternary value $t_i \in \{0,1,2\}$ cannot equal a function d_i of (essentially) uniformly distributed binary values $(b_{i_1},\dots,b_{i_q}) \in \{0,1\}^q$. Specifically, for a uniform $U \in \{0,1\}^q$ we have

$$\left| \Pr_{U \in \{0,1\}^q}[d_i(U) = 1] - \Pr[t_i = 1] \right| = \left| \frac{|\{x : d_i(x) = 1\}|}{2^q} - \frac{1}{3} \right| \geq 2^{-q}/3, \tag{5}$$

and thus, if we set the parameters so that $(b_{i_1}, \ldots, b_{i_q})$ and U are $(2^{-q}/3)$-close in statistical distance, Equations (4) and (5) give a contradiction. This proves the theorem in the case $Q_i \subseteq G$. The tradeoff between number of probes and redundancy in the conclusion of Theorem 1.1 corresponds, via the information-theoretic Lemma 2.2, to the tradeoff between the entropy of $Enc(t)$ in $\{0,1\}^u$ and the closeness of the random variable $b = Enc(t) \in \{0,1\}^u$ to uniform.

However, it may not be possible to find an index i such that $Q_i \subseteq G$. To circumvent this obstacle, we reason as follows. At the beginning of the argument, before applying the information-theoretic lemma, we identify a small set of *heavy* indexes $W \subseteq [u]$ that are in a noticeable fraction of the sets Q_i. We then fix the bits associated to the heavy indexes to their most likely value, so that we can still decode a large subset $T \subseteq \{0,1,2\}^n$ of arrays t. Then we seek an index i such that both the following hold for a uniformly selected $t \in T$: (I) the distribution of t_i is close to uniform over $\{0,1,2\}$, and (II) the distribution of $Enc(t)|_{Q_i}$ is close to uniform over $\{0,1\}^q$. Once we have such an index i, we obtain a contradiction by combining Equations (4) and (5), as explained above. To show the existence of such an index i we argue that most indexes satisfy (I) and also most indexes satisfy (II), which implies that some index will satisfy both. To show that most indexes satisfy (I), we again apply the information-theoretic lemma, using the fact that T is large in $\{0,1,2\}^n$. For (II), we also apply the information-theoretic lemma as explained earlier. However, we can now guarantee that most sets Q_i will not intersect the complement of G. This is because this complement is small and we have fixed all indexes that belong to noticeably many sets Q_i. This concludes the overview of the proof of Theorem 1.1, assuming that the probe locations are non-adaptive.

Comparisons with the arguments in [SV] and [Vio, Section 6.3]. The above argument is somewhat similar to one in [SV]. We mention the following alternative way to circumvent the obstacle that it may not be possible to find an index i such that $Q_i \subseteq G$. One can observe that every Q_i must intersect the complement of G. Since this complement is small, one can fix the associated bits so that we still decode a large subset $T \subseteq \{0,1,2\}^n$ of arrays t, but using at least one fewer probe. One can then repeat the argument q times to obtain a contradiction. The main results in this paper were initially obtained using this latter argument, which is similar to one in [Vio, Section 6.3]; but its inductive nature makes working out the parameters somewhat longer.

Handling adaptive probes. To explain how we handle adaptive probes, we note that the above argument holds for any decoder function d_i that satisfies Equation (5), and that adaptive decoders do satisfy Equation (5). Specifically, we model an adaptive decoder d_i by a decision tree of depth q in 2^q variables – thus we make the sets Q_i of probe locations exponentially bigger than in the non-adaptive case. It turns out we can afford this exponential increase in the size of the sets Q_i at no cost (roughly speaking, this is because we already needed statistical distance 2^{-q} even for the non-adaptive case). At the same time, each path of the decision tree is taken with probability 2^{-q}, and thus d_i still satisfies Equation (5). This concludes the overview of the proof of Theorem 1.1.

Since our arguments are similar to those in [SV] and [Vio, Section 6.3], one can ask why here we can handle adaptive probes, whereas the results in [SV, Vio] are stated for non-adaptive probes only. In fact, it can be shown that the results in [SV, Vio] hold for adaptive probes as well, but only when q is relatively small. This range of q is good enough for the results in this paper, some of which are in fact tight.

Extensions. It is clear at this point that the above argument applies to any other problem whose query answers have a probability mass function that is not a multiple of 2^{-q}. The dictionary problem (Theorem 1.2) is an example. Also, Theorems 1.1 and 1.2 continue to hold when replacing the constant 3 by any other constant which is not a power of 2.

Also, the probabilistic nature of our argument naturally applies to randomized representations, i.e., those that map any object t to a random string in $\{0,1\}^u$ that with high probability represents t: by an averaging argument one can fix the randomness to obtain a deterministic representation that works for a large subset T of objects, to which our technique applies. We note that [BMRV] considers a different kind of randomized data structure for representing sets, namely one in which the probes, not the encoding, is chosen at random. The technique in this paper might apply to that setting too, but we have not pursued this yet.

Finally, we mention that Theorem 1.1 immediately implies the same lower bound for the problem of representing $\{0,1,2\}^n$ in $\{0,1\}^n$ while answering *prefix sums* queries modulo 3, i.e., $S_t(i) := \sum_{j\le i} t_j$. Answering prefix sums in a group has been studied extensively; see, e.g., [Mil] — especially Open Problem 6 — and [PT, Păt]. Our lower bound applies here because one can reduce the problem of representing $\{0,1,2\}^n$ in $\{0,1\}^u$ to the prefix sum problem via the "telescoping" permutation

$$\pi(t_1, t_2, \ldots, t_n) := (t_1, -t_1 + t_2, -t_2 + t_3, \ldots, -t_{n-1} + t_n),$$

where all the arithmetic is modulo 3, which satisfies $S_{\pi(t)}(i) = t_i$ for every t and i.

2 Lower bound for representing ternary values in bits

In this section we prove our lower bound for representing ternary values in bits, i.e., Theorem 1.1. We start with a formal definition of the problem, and then we restate the theorem for the reader's convenience. The reader may want to consult Figure 1, which shows some of the relevant parameters, throughout this section.

We model an adaptive algorithm that decodes a ternary value t_i by a binary decision tree d_i of depth q. The internal nodes of the tree are labeled with one of 2^q binary variables, while the leaves are labeled with ternary values from $\{0,1,2\}$.

Definition 2.1 (Representing ternary values in bits)**.** *We say that we* represent $\{0,1,2\}^n$ in $\{0,1\}^u$ supporting single-element access by probing q bits *if there is a map* $Enc : \{0,1,2\}^n \to \{0,1\}^u$*,* n *sets* $Q_1, \ldots, Q_n \subseteq [u]$ *of size* 2^q *each, and* n *decision trees* $d_1, \ldots, d_n : \{0,1\}^{2^q} \to \{0,1,2\}$ *of depth* q *such that for every* $t \in \{0,1,2\}^n$ *and every* $i \in [n]$*:*

$$t_i = d_i\left(Enc(t)|_{Q_i}\right),$$

where $Enc(t)|_{Q_i}$ *denotes the* 2^q *bits of* $Enc(t) \in \{0,1\}^u$ *indexed by* Q_i*.*

Theorem 1.1 (Lower bound for representing ternary values)**.** (Restated.) *To represent* $\{0,1,2\}^n$ *in* $\{0,1\}^u$ *supporting single-element access by probing* q *bits, one needs*

$$u \geq (\log_2 3)n + n/2^{6q+22}.$$

In the rest of this section we prove Theorem 1.1. The proof makes use of the next lemma which was proved by Raz [Raz, Claim 5.1] and independently by Edmonds, Impagliazzo, Rudich, and Sgall [EIRS, Section 4]. The interested reader may also wish to look at Holenstein's formulation [Hol, Lemma 5]. We use here a version of the lemma which appears in [SV] and, unlike the above references, explicitly considers subsets of q random variables (see §A for an easy derivation of the next lemma from the results proved in [SV]). Before stating the lemma, let us recall some probability terminology. We say that two random variables V, W over the same set S are η-close if for every event $E \subseteq S$, $|\Pr[V \in E] - \Pr[W \in E]| \leq \eta$. Given a random variable V over a set S and an event E, we use $(V|E)$ to denote the probability distribution of V conditioned to E, that is for any event $A \subseteq E$, $\Pr_{(V|E)}[A] = \Pr[V \in A | V \in E]$.

Lemma 2.2 ([Raz, EIRS, SV])**.** *Let* $V = (V_1, \ldots, V_n)$ *be a collection of independent random variables where each one of them is distributed over a finite set* S *and equals any* $s \in S$ *with a probability that is a rational number. Let* $E \subseteq S^t$ *be an event such that* $\Pr[V \in E] \geq \epsilon$*. Then for any* $\eta > 0$ *and integer* q *there exists a set* $G \subseteq [n]$ *such that* $|G| \geq n - 16 \cdot q \cdot \log(1/\epsilon)/\eta^2$ *and for any* $i_1, \ldots, i_q \in G$ *the distributions*

$$(V_{i_1}, \ldots, V_{i_q} | V \in E) \quad \text{and} \quad (V_{i_1}, \ldots, V_{i_q})$$

are η*-close.*

2.1 Proof of Theorem 1.1

Let $u = (\log_2 3)n + r$, and assume for the sake of contradiction that

$$r < n/2^{6q+22}. \tag{6}$$

Definition 2.3. *A probe index* $j \in [u]$ *is* heavy *if* $\Pr_{i \in [n]}[j \in Q_i] \geq 1/\left(r \cdot 2^{3q+11}\right) =: \tau$*.*

Claim 2.3.1. *There are at most $2^q/\tau = 2^{4q+11} \cdot r$ heavy probes $j \in [u]$.*

Proof. For a fixed $j \in [u]$ and random $i \in [n]$ consider the indicator random variable $Y_j \in \{0,1\}$ that is 1 if and only if $j \in Q_i$. Then $2^q = E_{i\in[n]}[|Q_i|] = E_{i\in[n]}[\sum_{j\in[u]} Y_j] = \sum_{j\in[u]} \Pr_{i\in[n]}[j \in Q_i] \geq (\#\text{ heavy probes}) \cdot \tau$. □

Let $W \subseteq [u]$ be the set of

$$|W| \leq 2^{4q+11} \cdot r \tag{7}$$

heavy probes. The choice of $t \in \{0,1,2\}^n$ induces at most $2^{|W|}$ possibilities for the values $Enc(t)|_W$ of the heavy probes. Let $z \in \{0,1\}^{|W|}$ be the most common values for the heavy probes. By definition of z, there is a set $T \subseteq \{0,1,2\}^n$ of size

$$|T| \geq 3^n/2^{|W|} \tag{8}$$

such that for every $t \in T$ we have $Enc(t)|_W = z$; i.e., the values of the heavy probes for any $t \in T$ is fixed to z. Since these values are fixed, we can modify our decoding as follows. For every i define $Q'_i := Q_i \setminus W$ and also let d'_i be d_i where the values of the probes corresponding to variables in W have been fixed to the corresponding value in z. By renaming variables, letting $u' := u - |W|$ and $Enc' : \{0,1,2\}^n \to \{0,1\}^{u'}$ be Enc restricted to the bits in $[u] \setminus W$, we see that we are now representing T in $\{0,1\}^{u'}$ in the following sense: for every $t \in T$ and every $i \in [n]$:

$$t_i = d'_i\left(Enc'(t)|_{Q'_i}\right); \tag{9}$$

moreover, no probe $j \in [u']$ is heavy with respect to the sets Q'_i.

The next claim relies on our assumption (6) that we made for the sake of contradiction.

Claim 2.3.2. *There is an index $i \in [n]$ such that, for a randomly selected $t \in T$, both the following distributions are $(\eta := 1/2^{q+3})$-close to uniform: (I) the distribution $t_i \in \{0,1,2\}$, and (II) the distribution $Enc(t)|_{Q'_i} \in \{0,1\}^{2^q}$.*

Proof. We show that more than half the indexes $i \in [n]$ satisfy (I), and at least half the indexes $i \in [n]$ satisfy (II), which implies the existence of the desired index $i \in [n]$ that satisfies both (I) and (II).

(I): Consider choosing $t = (t_1, \ldots, t_n) \in \{0,1,2\}^n$ uniformly at random. Note that, using Inequality (8),

$$\Pr_t[t \in T] = \frac{|T|}{3^n} \geq \frac{1}{2^{|W|}}. \tag{10}$$

So by Lemma 2.2 (where the parameter q in the lemma is set to 1) and Inequalities (10, 7, 6) there is a set $H \subseteq [n]$ of size at least

$$|H| \geq n - 16 \cdot |W| \cdot 2^{2q+6} \geq n - 16 \cdot 2^{4q+11} \cdot r \cdot 2^{2q+6} = n - 2^{6q+21} \cdot r > n/2,$$

such that for every $i \in H$ the distribution $(t_i | t \in T)$ is (2^{-q-3})-close to uniform over $\{0,1,2\}$, i.e., it satisfies (I).

(II): Note that Enc is one-to-one – otherwise the hypothesis of the theorem is false – and so Enc' is also one-to-one by construction. Let $Enc'(T) := \{Enc'(t) : t \in T\}$. Consider choosing $b = (b_1, \ldots, b_{u'}) \in \{0,1\}^{u'}$ uniformly at random. Also using Inequality (8) and recalling that $u' = u - |W| = (\log_2 3)n + r - |W|$, we see that

$$\Pr_b[b \in Enc'(T)] = \frac{|Enc'(T)|}{2^{u'}} = \frac{|T|}{2^{u-|W|}} \geq \frac{3^n}{2^{|W|+u-|W|}} = \frac{3^n}{2^u} = \frac{1}{2^r}. \qquad (11)$$

Therefore, by Lemma 2.2 (where the parameter q in the lemma is set to the current 2^q) there is a set $G \subseteq [u']$ of size

$$|G| \geq u' - 16 \cdot 2^q \cdot r \cdot 2^{2q+6} = u' - 2^{3q+10} \cdot r$$

such that for any 2^q indexes $J = \{j_1, j_2, \ldots, j_{2^q}\} \subseteq G$, for a randomly selected $b \in \{0,1\}^{u'}$ the distribution $(b|_J | b \in Enc'(T))$ is (2^{-q-3})-close to random over $\{0,1\}^{2^q}$. Note that, because Enc' is one-to-one, the distribution $(b|_J | b \in Enc'(T))$ equals the distribution of $Enc'(t)|_J$ for a uniformly chosen $t \in T$. Thus, if $Q'_i \subseteq G$, we can set $J := Q'_i$ to see that the index i satisfies (II). To conclude, we make sure that $Q'_i \subseteq G$ for at least half of the indexes $i \in [n]$.

Let $\bar{G} := [u'] \setminus G$ denote the complement of G. Take a random $i \in [n]$. The probability that Q'_i intersects $\bar{G}$ is

$$\Pr_i[\exists j \in \bar{G} : j \in Q'_i] \leq \sum_{j \in \bar{G}} \Pr_i[j \in Q'_i] \leq |\bar{G}| \cdot \tau \leq 2^{3q+10} \cdot r \cdot \tau \leq 1/2.$$

In this last derivation we are using the union bound, then the fact that, after restricting to T, no probe index is heavy, and thus is in at most a $\tau = 1/\left(r \cdot 2^{3q+11}\right)$ fraction of the sets Q'_i (cf. Definition 2.3). Note here we crucially exploit the independence of the threshold τ for heaviness from the size of $\bar{G}$.

Therefore $Q'_i \subseteq G$ for at least half of the indexes $i \in [n]$; any such index satisfies (II). □

Claim 2.3.3. *The conclusion of Claim 2.3.2 is false.*

Proof. We show that the conclusion of Claim 2.3.2 leads to a contradiction. First, observe the following general fact: For any i and a uniformly distributed $U \in \{0,1\}^{2^q}$,

$$\left| \Pr_{U \in \{0,1\}^{2^q}}[d'_i(U) = 1] - 1/3 \right| \geq 2^{-q}/3. \qquad (12)$$

To see this we reason in two steps. First, let $X \subseteq \{0,1\}^q$ be the collection of paths in d'_i that lead from a root to a leaf that is labeled with 1. Since any path is taken with probability $1/2^q$ under U, we see that $\Pr_{U \in \{0,1\}^{2^q}}[d'_i(U) = 1] = |X|/2^q$. Note here we rely on the fact that the decision tree has depth q, although it is defined on 2^q variables. Second, if Equation (12) is false then $||X|/2^q - 1/3| < 2^{-q}/3$, which means $|3 \cdot |X| - 2^q| < 1$, and thus $3 \cdot |X| = 2^q$, which is impossible since 3 does not divide 2^q.

We now have:

$$
\begin{aligned}
2^{-q}/8 \geq & \left|\Pr_{t\in T}[t_i = 1] - 1/3\right| \qquad \text{(By (I) in the conclusion of Claim 2.3.2.)} \\
= & \left|\Pr_{t\in T}[d'_i\left(Enc'(t)|_{Q'_i}\right) = 1] - 1/3\right| \qquad \text{(By (9).)} \\
\geq & \left|\Pr_{U\in\{0,1\}^{2^q}}[d'_i\left(U\right) = 1] - 1/3\right| - 2^{-q}/8 \qquad \text{(By (II) in the conclusion of Claim 2.3.2.)} \\
\geq & \ 2^{-q}/3 - 2^{-q}/8, \qquad \text{(By (12).)}
\end{aligned}
$$

which is a contradiction. □

3 Lower bound for representing sets in bits

In this section we prove our lower bound for the dictionary problem, i.e., Theorem 1.2. The proof is very similar to that of Theorem 1.1. We start with a formal definition of the problem, and then we restate the theorem for the reader's convenience.

Definition 3.1 (Representing sets in bits). *We say that we* represent $S := \{x : x \in \{0,1\}^n, \sum_i x_i = n/3\}$ in $\{0,1\}^u$ answering membership queries by probing q bits *if there is a map $Enc : \{0,1\}^n \to \{0,1\}^u$, n sets $Q_1, \ldots, Q_n \subseteq [u]$ of size 2^q each, and n decision trees $d_1, \ldots, d_n : \{0,1\}^{2^q} \to \{0,1\}$ of depth q such that for every $t \in S$ and every $i \in [n]$:*

$$t_i = d_i\left(Enc(t)|_{Q_i}\right),$$

where $Enc(t)|_{Q_i}$ denotes the 2^q bits of $Enc(t) \in \{0,1\}^u$ indexed by Q_i.

Theorem 1.2 (Lower bound for representing sets). (Restated.) *For all sufficiently large n divisible by 3, to represent $S := \{x : x \in \{0,1\}^n, \sum_i x_i = n/3\}$ in $\{0,1\}^u$ answering membership queries by probing q bits, one needs*

$$u \geq \log_2 |S| + n/2^{6q+22} - \log_2 n.$$

3.1 Proof sketch of Theorem 1.2

We closely follow the proof of Theorem 1.1. Inequality (8) becomes

$$|T| \geq \frac{|S|}{2^{|W|}}.$$

We modify Claim 2.3.2 as follows:

Claim 3.1.1. *There is an index $i \in [n]$ such that, for a randomly selected $t \in T$, both the following are true: (I) the distribution $t_i \in \{0,1\}$ is $(1/2^{q+3})$-close to the distribution that puts weight 1/3 on 1, and (II) the distribution $Enc(t)|_{Q'_i}$ is $(1/2^{q+3})$-close to uniform over $\{0,1\}^{2^q}$.*

Proof sketch of Claim 3.1.1. The argument for (II) is identical to that of Claim 2.3.2.

The argument for (I) is modified as follows. We choose $t = (t_1, \ldots, t_n) \in \{0,1\}^n$ where the binary variables t_i are independent and take value 1 with probability 1/3. Using that each element in T has weight $n/3$, and then standard estimates [CT, Lemma 17.5.1], we obtain

$$\Pr[t \in T] \geq \frac{|T|}{2^{H(1/3)n}} \geq \frac{|T|}{|S| \cdot \Theta(\sqrt{n})} \geq \frac{1}{2^{|W|} \cdot n}$$

for sufficiently large n. The application of Lemma 2.2 now yields a set H of size

$$n - 2^{2q+10}(|W| + \log n) \geq n - 2^{6q+21}(r + \log n)$$

which is strictly bigger than $n/2$ under the assumption that $r < n/2^{6q+22} - \log n$. □

4 Logarithmic forms and cell probes

In this section we highlight a link between number theory and the problem of representing ternary values in bits; we then discuss the relevance of this link to the challenge of proving lower bounds in the cell-probe model. Let us start by recalling from §1.1 a simple block-wise approach to represent an array of n ternary values $t = (t_1, \ldots, t_n) \in \{0,1,2\}^n$ in terms of u bits $b \in \{0,1\}^u$: We use arithmetic coding for each block of k ternary values; to access a value $t_i \in \{0,1,2\}$, we probe the $q := \lceil (\log_2 3)k \rceil$ bits of the encoding of the block containing it. The space used is

$$u = \lceil (\log_2 3)k \rceil \cdot n/k = (\log_2 3)n + \epsilon \cdot n/k, \tag{13}$$

where

$$\epsilon := \lceil (\log_2 3)k \rceil - (\log_2 3)k > 0 \tag{14}$$

is the distance of $(\log_2 3)k$ from the next integer.

We note that any lower bound on the redundancy of representations of ternary values in bits implies a corresponding lower bound on ϵ; for example, our Theorem 1.1 implies a lower bound of the form $\epsilon \geq 1/2^{O(k)}$. We also note that lower bounds on ϵ depending on k are related to well-studied questions in number theory. In particular, the results on logarithmic forms by A. Baker and N. I. Feldman, a special case of which is stated next, imply the stronger bound $\epsilon \geq 1/k^{O(1)}$.

Theorem 4.1 (Theorem 3.1 in [Bak]). *There is an absolute constant $c > 0$ such that for all positive integers k and ℓ we have*

$$|\ell \log_e 2 - k \log_e 3| \geq 1/(\max\{\ell, k, 2\})^c.$$

In particular, there is an absolute constant $c > 0$ such that for every integer $k \geq 0$ and every integer $\ell \geq c$ we have

$$|\ell - k \log_2 3| \geq 1/\ell^c.$$

Remark on the proof of Theorem 4.1. The proof of a generalization of the first claim in the statement of Theorem 4.1 can be found in [Bak, §3], while a more recent account of the subject is in [BW].

The "in particular" part is obtained as follows. We can assume without loss of generality that $k \leq \ell$, for else for sufficiently large ℓ we have $|\ell - k\log_2 3| \gg 1 \geq 1/\ell$ and the theorem is proved. Dividing the inequality in the first part of the theorem by $\log_e 2 \in (0,1)$ we then have

$$|\ell - k\log_2 3| \geq (\log_2 e)/(\max\{\ell, k, 2\})^c \geq 1/\ell^c.$$

□

We now discuss the relevance of Theorem 4.1 to the cell-probe model, where recall the u bits of memory are divided in cells of $\log n$ bits, and each probe returns the content of an entire cell. In this model, Pătraşcu [Păt] gives for every constant $c \geq 0$ a representation that uses space

$$u \leq (\log_2 3)n + n/\log^c n \tag{15}$$

and where we access each ternary value t_i by probing a number of cells that depends on c only. Since the block-wise representation recalled at the beginning of this section is immediately implementable in the cell-probe model, and there we access each t_i by probing just 1 cell of $\log n := \lceil(\log_2 3)k\rceil$ bits, we see from the expression for its space (13) that to prove a lower bound that matches (15) some form of Theorem 4.1 is necessary. We observe next that Theorem 4.1 is sufficient to obtain the following lower bound for one cell probe.

Theorem 4.2. *Let* $\log_2 n$ *and* $u := n/\log_2 n$ *be sufficiently large integers. To represent* $\{0,1,2\}^n$ *in* $\{0,1\}^u$ *supporting single-element access by probing* 1 *cell of* $\log n$ *bits, one needs*

$$u \geq (\log_2 3)n + n/\log^{O(1)} n.$$

Proof. Let $n = 2^\ell$. Let k be the maximum over all cells i of the number of ternary values that probe i. Note that $3^k \leq n$, and so $k\log_2 3 \leq \ell$. By Theorem 4.1, there is an absolute constant $c > 0$ such that $k\log_2 3 \leq \ell - 1/\ell^c$. Since each of the n ternary values must probe one of the u/ℓ cells, we have

$$n \leq \frac{u}{\ell}k \leq \frac{u}{\ell \cdot \log_2 3}\left(\ell - \frac{1}{\ell^c}\right) = \frac{u}{\log_2 3}\left(1 - \frac{1}{\ell^{c+1}}\right).$$

Since $(1-\alpha)^{-1} \geq 1 + \alpha$ for every $\alpha \in [0,1)$, we obtain

$$u \geq n\log_2 3\left(1 + \frac{1}{\ell^{c+1}}\right).$$

□

Open problem: Extend Theorem 4.2 to hold for two cell probes.

Acknowledgments. We thank Ronen Shaltiel for useful discussions.

References

[Bak] A. Baker. *Transcendental number theory.* Cambridge Mathematical Library. Cambridge University Press, Cambridge, second edition, 1990. 10, 11

[BW] A. Baker and G. Wüstholz. *Logarithmic forms and Diophantine geometry*, volume 9 of *New Mathematical Monographs.* Cambridge University Press, Cambridge, 2007. 11

[BMRV] H. Buhrman, P. B. Miltersen, J. Radhakrishnan, and S. Venkatesh. Are bitvectors optimal? *SIAM J. Comput.*, 31(6):1723–1744 (electronic), 2002. 2, 5

[CT] T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing).* Wiley-Interscience, 2006. 10

[EIRS] J. Edmonds, R. Impagliazzo, S. Rudich, and J. Sgall. Communication complexity towards lower bounds on circuit depth. *Computational Complexity*, 10(3):210–246, 2001. 3, 6

[GM] A. Gál and P. B. Miltersen. The cell probe complexity of succinct data structures. *Theoret. Comput. Sci.*, 379(3):405–417, 2007. 1

[GRRR] R. F. Geary, N. Rahman, R. Raman, and V. Raman. A simple optimal representation for balanced parentheses. *Theor. Comput. Sci.*, 368(3):231–246, 2006. 1

[GGG$^+$] A. Golynski, R. Grossi, A. Gupta, R. Raman, and S. S. Rao. On the Size of Succinct Indices. In L. Arge, M. Hoffmann, and E. Welzl, editors, *ESA*, volume 4698 of *Lecture Notes in Computer Science*, pages 371–382. Springer, 2007. 1

[GRR] A. Golynski, R. Raman, and S. S. Rao. On the Redundancy of Succinct Data Structures. In J. Gudmundsson, editor, *SWAT*, volume 5124 of *Lecture Notes in Computer Science*, pages 148–159. Springer, 2008. 1

[Hol] T. Holenstein. Parallel repetition: simplifications and the no-signaling case (extended abstract). In *STOC'07—Proceedings of the 39th Annual ACM Symposium on Theory of Computing*, pages 411–419. ACM, New York, 2007. 6

[Mil] P. B. Miltersen. Cell probe complexity - a survey. In *In 19th Conference on the Foundations of Software Technology and Theoretical Computer Science (FSTTCS), 1999. Advances in Data Structures Workshop*, 1999. 2, 5

[MP] M. Minsky and S. Papert. *Perceptrons.* MIT Press, Cambridge, MA, 1969. 2

[Mit] M. Mitzenmacher. Compressed bloom filters. In *PODC '01: Proceedings of the twentieth annual ACM symposium on Principles of distributed computing*, pages 144–150, New York, NY, USA, 2001. ACM. 1

[Pag] R. Pagh. Low redundancy in static dictionaries with constant query time. *SIAM J. Comput.*, 31(2):353–363 (electronic), 2001. 1, 2

[Pǎt] M. Pǎtraşcu. Succincter. In *49th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2008. 1, 2, 5, 11

[PT] M. Pǎtraşcu and C. E. Tarniţǎ. On dynamic bit-probe complexity. *Theoret. Comput. Sci.*, 380(1-2):127–142, 2007. 5

[Raz] R. Raz. A parallel repetition theorem. *SIAM J. Comput.*, 27(3):763–803 (electronic), 1998. 3, 6

[SV] R. Shaltiel and E. Viola. Hardness amplification proofs require majority. In *Proceedings of the 40th Annual ACM Symposium on the Theory of Computing (STOC)*, Victoria, Canada, 17–20 May 2008. 3, 4, 5, 6, 13

[Vio] E. Viola. *The Complexity of Hardness Amplification and Derandomization*. PhD thesis, Harvard University, 2006. `http://www.eccc.uni-trier.de/eccc`. 4, 5

A Proof of Lemma 2.2

Proof of Lemma 2.2. [SV] gives a proof for the case that the variables V_i are uniformly distributed. We now explain how to handle the more general case in which they are not uniformly distributed; this uses a standard trick which also appears in [SV]. Model the variables $(V_1, \ldots, V_n)$ by independent variables $(W_1, \ldots, W_n)$ uniformly distributed over a set S' and a function $f : S' \to S$ so that $f(W_i)$ is distributed like V_i. This is possible for a sufficiently large S' because of our assumption that each variable V_i equals any $s \in S$ with a probability that is a rational number. Now apply the lemma to the uniformly distributed $(W_1, \ldots, W_n)$ with respect to the event that $(f(W_1), \ldots, f(W_n)) \in A$. This gives a set $G \subseteq [n]$ such that $|G| \geq n - 16 \cdot q \cdot a/\eta^2$ and for any $i_1, \ldots, i_q \in G$ the distributions

$$(W_{i_1}, \ldots, W_{i_q} | (f(W_1), \ldots, f(W_n)) \in A), \quad \text{and} \quad (W_{i_1}, \ldots, W_{i_q})$$

are η-close. This implies that the distributions

$$(f(W_{i_1}), \ldots, f(W_{i_q}) | (f(W_1), \ldots, f(W_n)) \in A) = (V_{i_1}, \ldots, V_{i_q} | (V_1, \ldots, V_n) \in A) \quad \text{and}$$
$$(f(W_{i_1}), \ldots, f(W_{i_q})) = (V_{i_1}, \ldots, V_{i_q})$$

are η-close, as desired. □