On Regularity-Preserving Functions

Dexter Kozen
Computer Science Department
Cornell University
Ithaca, NY 14853-7501, USA
kozen@cs.cornell.edu

1 Some Homework Exercises

In introductory automata theory, one can find a wealth of entertaining automata-theoretic puzzles such as the classical *first halves problem*:

> Show that if A is a regular set, then so is the set of all first halves of strings in A:
>
> $$FirstHalves(A) = \{x \mid \exists y\ |y| = |x| \text{ and } xy \in A\} .$$

Once students master the basic pebbling technique for solving such problems, they can move on to more challenging variants:

> Show that if A is a regular set, then so are the following:
>
> $$\begin{aligned} A_{n^2} &= \{x \mid \exists y\ |y| = |x|^2 \text{ and } xy \in A\} \\ A_{2^n} &= \{x \mid \exists y\ |y| = 2^{|x|} \text{ and } xy \in A\} \\ A_{2^{2^n}} &= \{x \mid \exists y\ |y| = 2^{2^{|x|}} \text{ and } xy \in A\} . \end{aligned}$$

Students are often quite surprised at first that these sets should be regular, since the presence of the nonlinear functions seems to contradict their emerging intuition about regularity.

An effective tool in all these problems is the *Boolean transition matrix* of an automaton for A. This is the square Boolean matrix Δ indexed by states of the automaton with 1 in position uv iff the automaton contains a transition $u \xrightarrow{a} v$ for some symbol a. The Boolean powers Δ^n give the n-step transition relations. To solve the problem above for A_{2^n}, for example, one only has to determine how to get from Δ^{2^n} to $\Delta^{2^{n+1}}$ in one step. This is done by squaring the matrix. There are only finitely many possible such matrices, so they can all be encoded in the finite control of an automaton for A_{2^n}. Composing this construction with itself gives an automaton for $A_{2^{2^n}}$.

Similarly, to solve the problem for A_{n^2}, one has to determine how to get from Δ^{n^2} to $\Delta^{(n+1)^2}$ in one step. Observing that $\Delta^{(n+1)^2} = \Delta^{n^2}\Delta^{2n}\Delta$, we see that it would also be nice to know Δ^{2n}. But this is no problem, since Δ^{2n} can be maintained in the state as well. Thus the states of the new automaton encode pairs (C, D) of matrices, along with transitions $(C, D) \longrightarrow (CD\Delta, D\Delta^2)$. One can then prove easily by induction that in n steps, $(I, I) \xrightarrow{n} (\Delta^{n^2}, \Delta^{2n})$.

Expanding on this idea leads to an elegant proof that if A is regular, then so is the set

$$A_p \quad = \quad \{x \mid \exists y \ |y| = p(|x|) \text{ and } xy \in A\} \ ,$$

where p is any polynomial with nonnegative integer coefficients. The solution is based on the hint

$$p(n+1) \quad = \quad \sum_{i=0}^{d} \frac{p^{(i)}(n)}{i!} \tag{1}$$

where d is the degree of p and $p^{(i)}$ is the i^{th} derivative. From (1) we have

$$p^{(j)}(n+1) \quad = \quad \sum_{i=0}^{d-j} \frac{p^{(i+j)}(n)}{i!} \quad = \quad \sum_{k=j}^{d} \frac{p^{(k)}(n)}{(k-j)!} \ ,$$

therefore

$$\frac{p^{(j)}(n+1)}{j!} \quad = \quad \sum_{k=j}^{d} \left(\begin{array}{c} k \\ j \end{array} \right) \frac{p^{(k)}(n)}{k!} \ . \tag{2}$$

Also,

$$\frac{p^{(j)}(0)}{j!} \quad = \quad a_j \ , \tag{3}$$

where a_j is the coefficient of n^j in $p(n)$.

Now one builds an automaton for A_p whose states encode $(d+1)$-tuples of matrices

$$(C_i \mid 0 \le i \le d) \quad = \quad (C_0, C_1, \ldots, C_d)$$

and transitions

$$(C_i \mid 0 \le i \le d) \quad \longrightarrow \quad (\prod_{k=j}^{d} C_k^{\binom{k}{j}} \mid 0 \le j \le d) \ .$$

One can then show by induction using (2) and (3) that in n steps,

$$(\Delta^{a_i} \mid 0 \le i \le d) \quad \xrightarrow{n} \quad (\Delta^{p^{(i)}(n)/i!} \mid 0 \le i \le d) \ .$$

In particular, the first component is $\Delta^{p(n)}$.

2 Regularity-Preserving Functions

Exercises like these arouse one's curiosity about the general class of functions f for which the theorem

If A is regular, then so is

$$A_f \quad = \quad \{x \mid \exists y \ |y| = f(|x|) \text{ and } xy \in A\}$$

holds. Does this class have a nice characterization? Let us call such functions *regularity preserving.* Not all functions are regularity preserving: for example, $\log n$ is not. The class is closed under addition, multiplication, exponentiation, composition, and contains arbitrarily fast growing functions, including highly noncomputable ones.

In the remainder of this note we give two characterizations of the class of regularity-preserving functions in terms of the concept of *ultimate periodicity.* One of these characterizations involves two simple independent conditions that are relatively easy to check.

Let Σ be a finite alphabet, Σ^* the set of finite-length strings over Σ, $\mathbb{N} = \{0, 1, 2, \ldots\}$. Subsets of Σ^* are denoted $A, B, \ldots$ and subsets of $\mathbb{N}$ are denoted $U, V, \ldots$ The length of a string $x \in \Sigma^*$ is denoted $|x|$. The set of all lengths of strings in A is denoted $\text{lengths}(A)$.

Definition 1 A set $U \subseteq \mathbb{N}$ is called *ultimately periodic (u.p.)* (or *semilinear*) if

$$\exists p \geq 1 \ \overset{\infty}{\forall} n \quad n \in U \leftrightarrow n + p \in U \ .$$

More generally, a function $f : \mathbb{N} \to \mathbb{N}$ is called *ultimately periodic* if

$$\exists p \geq 1 \ \overset{\infty}{\forall} n \quad f(n) = f(n + p) \ .$$

□

Here $\overset{\infty}{\forall}$ means "for all but finitely many." Note that a set is u.p. iff its characteristic function is. The number p is called a *period* of U or f. Every u.p. set or function has a smallest period, which is the gcd of all its periods.

A simple example of a u.p. set is $[k]_m$, the congruence class of k modulo m; *i.e.*, the set of numbers n such that $m \mid n - k$. In fact, the family of u.p. sets is the smallest family containing all finite sets and the sets $[k]_m$ and closed under the Boolean operations. If U, V are u.p. with periods p, q respectively, then $U \cup V$ is u.p. with period $\mathrm{lcm}(p, q)$.

It is well known (and not difficult to prove) that for any regular set A, the set $\mathrm{lengths}(A)$ is u.p.; and for any u.p. set U, the set $\{x \mid |x| \in U\}$ is regular. In particular, if A a set of strings over a single letter alphabet, then A is regular iff $\mathrm{lengths}(A)$ is u.p.

Definition 2 A function $f : \mathbb{N} \to \mathbb{N}$ is said to *preserve ultimate periodicity* if $f^{-1}(U)$ is u.p. whenever U is. □

Definition 3 A function $f : \mathbb{N} \to \mathbb{N}$ is said to be *ultimately periodic modulo m (u.p.* mod *m)* if the function $n \mapsto f(n) \bmod m$ is ultimately periodic. □

For $A \subseteq \Sigma^*$ and $f : \mathbb{N} \to \mathbb{N}$, let

$$\begin{aligned} A_f &= \{x \mid \exists y \ |y| = f(|x|) \text{ and } xy \in A\} \\ A'_f &= \{x \mid \exists y \ |y| = f(|x|) \text{ and } y \in A\} \ . \end{aligned}$$

Consider the following four conditions:

C1 A_f is regular whenever A is.

C2 A'_f is regular whenever A is.

C3 f preserves ultimate periodicity.

C4 (i) f is ultimately periodic modulo m for all $m \geq 1$; and
(ii) $f^{-1}(\{x\})$ is ultimately periodic for all $x \in \mathbb{N}$.

We remark that the two subconditions of **C4** are independent. For any $U \subseteq \mathbb{N}$, the function

$$g(n) \quad = \quad \begin{cases} 0\ , & \text{if } n \in U, \\ n!\ , & \text{if } n \notin U \end{cases}$$

satisfies **C4**(i), since for any $n \geq m$, $g(n) = 0 \bmod m$; but $g^{-1}(\{0\}) = U$, so **C4**(ii) fails when U is not u.p. On the other hand, the function

$$\begin{aligned} h(n) \quad &= \quad k, \text{ where } 2^k \text{ is the highest power of 2 dividing } n+1 \\ &= \quad \text{the position of the first 0 in the binary representation of } n, \\ &\qquad \text{reading from right to left} \end{aligned}$$

satisfies **C4**(ii), since $h^{-1}(\{k\}) = [2^k - 1]_{2^{k+1}}$, but not **C4**(i), since for any $p \geq 1$ there are arbitrarily large n such that $h(n)$ is odd iff $h(n+p)$ is even: if $k = h(p-1)+2$, then for any $m \geq 1$,

$$\begin{aligned} h(2^{k^m+1} - 1) \quad &= \quad k^m + 1 \\ h(2^{k^m+1} - 1 + p) \quad &= \quad k - 2\ . \end{aligned}$$

Thus h is not u.p. modulo 2.

Lemma 4 *The statement* **C4***(i) is equivalent to the statement that* $f^{-1}([i]_m)$ *is ultimately periodic for all* i *and* m*.*

Proof. For all m,

$$\begin{aligned} & f^{-1}([i]_m) \text{ is u.p., } 0 \leq i \leq m-1 \\ \leftrightarrow \quad & \bigwedge_{i=0}^{m-1} \exists p_i \geq 1\ f^{-1}([i]_m) \text{ is u.p. with period } p_i \end{aligned}$$

$$
\begin{aligned}
&\leftrightarrow \quad \exists p \geq 1 \ \bigwedge_{i=0}^{m-1} f^{-1}([i]_m) \text{ is u.p. with period } p \text{ (take } p = \mathrm{lcm}_i\ p_i)\\
&\leftrightarrow \quad \exists p \geq 1 \ \bigwedge_{i=0}^{m-1} \overset{\infty}{\forall} n \quad n \in f^{-1}([i]_m) \leftrightarrow n + p \in f^{-1}([i]_m)\\
&\leftrightarrow \quad \exists p \geq 1 \ \bigwedge_{i=0}^{m-1} \overset{\infty}{\forall} n \quad f(n) \in [i]_m \leftrightarrow f(n+p) \in [i]_m\\
&\leftrightarrow \quad \exists p \geq 1 \ \overset{\infty}{\forall} n \ \bigwedge_{i=0}^{m-1} \quad f(n) \in [i]_m \leftrightarrow f(n+p) \in [i]_m\\
&\leftrightarrow \quad \exists p \geq 1 \ \overset{\infty}{\forall} n \quad f(n) = f(n+p) \bmod m\\
&\leftrightarrow \quad f \text{ is u.p. modulo } m.
\end{aligned}
$$

□

Theorem 5 *The four conditions* **C1** – **C4** *are equivalent.*

Proof. (**C1** $\to$ **C4**) To show **C4**(i), let $0 \leq k \leq m-1$, and consider the regular set $(a^m)^* a^k$. We have

$$
\begin{aligned}
((a^m)^* a^k)_f &= \{x \mid \exists y \ |y| = f(|x|) \text{ and } xy \in \{a^{mn+k} \mid n \geq 0\}\}\\
&= \{a^i \mid \exists j \ j = f(i) \text{ and } a^i a^j \in \{a^{mn+k} \mid n \geq 0\}\}\\
&= \{a^i \mid \exists j \ j = f(i) \text{ and } i + j = k \bmod m\}\\
&= \{a^i \mid i + f(i) = k \bmod m\}\ ,
\end{aligned}
$$

and by **C1**, this set is regular, thus

$$
\begin{aligned}
\mathrm{lengths}(((a^m)^* a^k)_f) &= \mathrm{lengths}(\{a^i \mid i + f(i) = k \bmod m\})\\
&= \{i \mid i + f(i) = k \bmod m\}\\
&= f'^{-1}([k]_m)
\end{aligned}
$$

is u.p., where $f'(n) = n + f(n)$. Since this holds for arbitrary k and m, it follows from Lemma 4 that f' satisfies **C4**(i), thus $f'(n)$ is u.p. modulo m for any m. Since the function $n \mapsto (-n) \bmod m$ is also u.p., so is the sum

$$
\begin{aligned}
f'(n) \bmod m + (-n) \bmod m &= f'(n) - n \bmod m\\
&= f(n) \bmod m\ .
\end{aligned}
$$

To show **C4**(ii), consider the regular set a^*ba^k. Intersecting $(a^*ba^k)_f$ with the regular set a^*b, we obtain

$$\begin{aligned}
&a^*b \cap (a^*ba^k)_f \\
&= \{a^nb \mid \exists y\ |y| = f(|a^nb|) \text{ and } a^nby \in \{a^nba^k \mid n \geq 0\}\} \\
&= \{a^nb \mid \exists y\ |y| = f(n+1) \text{ and } y = a^k\} \\
&= \{a^nb \mid k = f(n+1)\} \\
&= \{a^nb \mid n+1 \in f^{-1}(\{k\})\}\ ,
\end{aligned}$$

and by **C1** this set is regular, therefore

$$\begin{aligned}
\text{lengths}(\{a^nb \mid n+1 \in f^{-1}(\{k\})\}) &= \{n+1 \mid n+1 \in f^{-1}(\{k\})\} \\
&= f^{-1}(\{k\}) - \{0\}
\end{aligned}$$

is u.p. Then $f^{-1}(\{k\})$ is u.p. as well.

(**C4** $\rightarrow$ **C3**) Let U be a u.p. set with period p. Then U can be expressed as a Boolean combination of a finite set F and sets of the form $[i]_p$:

$$U = F \oplus ([i_1]_p \cup [i_2]_p \cup \cdots \cup [i_k]_p)\ ,$$

where $\oplus$ denotes symmetric difference of sets. Then

$$\begin{aligned}
&f^{-1}(U) \\
&= f^{-1}(F \oplus ([i_1]_p \cup [i_2]_p \cup \cdots \cup [i_k]_p)) \\
&= f^{-1}(F) \oplus (f^{-1}([i_1]_p) \cup f^{-1}([i_2]_p) \cup \cdots \cup f^{-1}([i_k]_p)) \\
&= (\bigcup_{x \in F} f^{-1}(\{x\})) \oplus (f^{-1}([i_1]_p) \cup f^{-1}([i_2]_p) \cup \cdots \cup f^{-1}([i_k]_p))\ .
\end{aligned}$$

By **C4**, Lemma 4, and the closure of u.p. sets under the Boolean operations, this set is u.p.

(**C3** $\rightarrow$ **C2**) Note that

$$\begin{aligned}
A'_f &= \{x \mid \exists y \in A\ |y| = f(|x|)\} \\
&= \{x \mid \exists n \in \text{lengths}(A)\ n = f(|x|)\} \\
&= \{x \mid f(|x|) \in \text{lengths}(A)\} \\
&= \{x \mid |x| \in f^{-1}(\text{lengths}(A))\}\ .
\end{aligned}$$

If A is regular, then lengths(A) is u.p. By **C3**, $f^{-1}(\text{lengths}(A))$ is u.p., therefore A'_f is regular.

$(\mathbf{C2} \to \mathbf{C1})$ In the notation of [1], let A be any regular set and let $M = (Q,\ \Sigma,\ \delta,\ s,\ F)$ be a deterministic finite automaton with $L(M) = A$. If $p \in Q$ and $G \subseteq Q$, let M_p^G be the automaton

$$M_p^G \;=\; (Q,\ \Sigma,\ \delta,\ p,\ G)\ .$$

Then

$$\begin{aligned}
A_f &= \{x \mid \exists y\ |y| = f(|x|) \text{ and } xy \in A\} \\
&= \{x \mid \exists y\ |y| = f(|x|) \text{ and } \delta(s, xy) \in F\} \\
&= \{x \mid \exists y\ |y| = f(|x|) \text{ and } \delta(\delta(s,x), y) \in F\} \\
&= \bigcup_{p \in Q} \{x \mid \exists y\ |y| = f(|x|) \text{ and } \delta(s,x) = p \text{ and } \delta(p,y) \in F\} \\
&= \bigcup_{p \in Q} \{x \mid \delta(s,x) = p\} \cap \{x \mid \exists y\ |y| = f(|x|) \text{ and } \delta(p,y) \in F\} \\
&= \bigcup_{p \in Q} L(M_s^{\{p\}}) \cap L(M_p^F)'_f\ .
\end{aligned}$$

By **C2** and the closure of the regular sets under the Boolean set operations, this is a regular set. □

It follows from the various characterizations of Theorem 5 that the regularity-preserving functions are closed under addition, multiplication, exponentiation, and composition. The function $\log n$ is not regularity preserving, because it is not ultimately periodic modulo 2.

Acknowledgement

Devdatt Dubhashi first observed the equivalence of **C1** and **C3**.

References

[1] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages and Computation.* Addison-Wesley, 1979.