Borel Normality, Automata, and Complexity

Wolfgang Merkle and Jan Reimann

Institut für Informatik
Universität Heidelberg
The Quest for Randomness

• Intuition: An infinite sequence of fair coin tosses (H/T) will produce

\[ H \text{ with an asymptotic frequency of } \frac{1}{2}. \]  

(*)

• Measure Theory: The law of large numbers asserts the set of sequences satisfying (*) has measure one with respect to the uniform Bernoulli measure (Lebesgue measure).

• Collectives: Von Mises tried to base probability on individual objects. Probabilities could be assigned by studying a single instance in a Collective (Kollektiv).

(“First the collective, then the probability.”)
Von Mises’ Collectives

Von Mises gave two ’axioms’ for collectives:

1. The asymptotic frequency of occurrences of H in the collective equals 1/2.

2. Property (1) persists for any subsequence of outcomes derived from the collective by an admissible place selection rule.

Problem: What is an admissible selection rule?

- Admissible: Select all even/odd/prime/... positions.
- Not admissible: Given a sequence H T H T T H..., select all positions where H occurs.
Selection Rules

How to select a subsequence from a given sequence $A \in \{0, 1\}^\infty$?

- **Oblivious selection rule**: sequence $S \in \{0, 1\}^\infty$. Subsequence $B = A / S$ obtained: all the bits $A(i)$ with $S(i) = 1$.

- **(General) Selection rule**: language $L \subseteq \{0, 1\}^*$. Subsequence $B = A / L$ obtained: the bits $A(i)$ such that the prefix $A(0) \ldots A(i-1)$ is in $L$. 

VIC 2004 – p.4/26
Stochasticity

- Church proposed to admit only computable selection rules.
- This lead to the study of stochastic sequences. (Church, Wald, Kolmogorov, Loveland, . . .)
- Definition: A sequence $S \in \{0, 1\}^\infty$ is (Church-) stochastic if
  $$\lim_{n \to \infty} \frac{\#_1(A/L \upharpoonright n)}{n} = \frac{1}{2}$$
  for any computable language $L$.
- Note that $L = \{0, 1\}^*$ is admissible, hence every stochastic sequence has limiting frequency $1/2$. 
Normal Sequences

A sequence \( N \in \{0, 1\}^\infty \) is normal if any word \( w \) of length \( n \) appears as a subword of \( N \) with frequency \( 2^{-n} \).

More formally, for every \( w \in \{0, 1\}^k \), it holds that

\[
\lim_{n \to \infty} \frac{\#_w(N \upharpoonright n)}{n} = \frac{1}{2^{-k}}
\]

where

\[
\frac{\#_w(N \upharpoonright n)}{n} \overset{\text{def}}{=} \frac{|\{i \leq n - k : N \upharpoonright i...i+k-1 = w\}|}{n}.
\]
Facts about Normality

• **Borel**: Almost every sequence is normal (with respect to Lebesgue measure).

• Normality is not base-invariant (**Cassels**).

• Few explicit normal sequences are known:
  ◦ **Champernowne** (base 10): 1234567891011121314…
  ◦ **Copeland-Erdös** (base 10): 23571113171923293137…

• Many open questions, e.g.: Is $\pi$ normal?
Normal Sequences as Collectives

- **Obvious:** Not all normal sequences are stochastic. (Can be algorithmically quite easy, e.g. Champernowne's sequence)

- **Question:** Which selection rules do preserve normality?

- For **oblivious selection rules:** Kamae gave a complete characterization in terms of measures generated by sequences under shift map.
Oblivious Selection Rules

• Let $T$ be the **shift map**, transforming a sequence $A = A(0)A(1)A(2)\ldots$ into another sequence by cutting off the first bit, i.e. $T(A) = A(1)A(2)A(3)\ldots$.

• Given a sequence $A$, $\delta_A$ denotes the **Dirac measure** induced by $A$, that is, for any set $B$ of sequences,

$$
\delta_A(B) = \begin{cases} 
1 & \text{if } A \in B, \\
0 & \text{otherwise}. 
\end{cases}
$$
Theorem: [Kamae, 1973] An oblivious selection rule $S$ preserves normality if and only if $S$ is completely deterministic, that is, any cluster point (in the weak topology) of the measures $\mu_n = \frac{1}{n} \sum_{i=0}^{n-1} \delta_{T^i(s)}$ has entropy 0.

Note that if a sequence $A$ is normal, then any cluster point of the measures $\mu_n$ is the uniform $(1/2, 1/2)$-Bernoulli measure, which has entropy 1.
Oblivious Selection Rules

• **Example** of a completely deterministic sequence: For any real $\alpha > 1$, take the characteristic sequence of the set

$$\{[j\alpha] : j \geq 1\}.$$ 

• It follows that there are uncountably many completely deterministic sequences, hence there are many that are quite complicated, from an algorithmic point of view.

• **Sturmian trajectories**: Symbolic coding of irrational rotations of the circle.

• **Theorem**: Every Turing degree contains a Sturmian trajectory.
Normality and Finite Automata

- For general selection rules: Fundamental result by Agafonoff [1968], Schnorr and Stimm [1972], and Kamae and Weiss [1975].

- **Theorem**: If L is regular, then L preserves normality.

- More automata-theoretic style proofs were given by O’Connor [1988] and Broglio and Liardet [1992]

- Uses an ergodic feature of finite automata.
Kamae and Weiss [1975] asked if normality is preserved by larger classes of languages, too (e.g. context-free languages).

**Answer:** If larger, then not much!

By varying Champernowne's construction, we give two counterexamples:

1. A normal sequence not preserved by a deterministic one-counter language (accepted by a deterministic pushdown automata with unary stack alphabet).
2. A normal sequence that is not preserved by a linear language (slightly more complicated).
Theorem: There exists a deterministic one-counter language $L$ and a normal sequence $\tilde{N}$ such that the sequence $\tilde{N}/_L$ selected from $\tilde{N}$ by $L$ is infinite and constant.
Constructing $\tilde{N}$

- For any $n$, let

$$v_n = 0^n 0^{n-1}1 0^{n-2}10 \ldots 1^n$$

be the word that is obtained by concatenating all words of length $n$ in lexicographic order.

- **Definition:** A set $W \subseteq \{0, 1\}^*$ of words is normal in the limit if for any nonempty word $u$ and any $\varepsilon > 0$ for all but finitely many words $w$ in $W$,

$$\frac{1}{2|u|} - \varepsilon < \frac{\#_u(w)}{|w|} < \frac{1}{2|u|} + \varepsilon.$$
Constructing $\tilde{N}$

- **Proposition**: The set $\{v_1, v_2, \ldots\}$ is normal in the limit.
- **Lemma**: [Champernowne] Let $W$ be a set of words that is normal in the limit. Let $w_1, w_2, \ldots$ be a sequence of words in $W$ such that

\[
\forall w \in W \frac{|\{i \leq t: w_i = w\}|}{t} \xrightarrow{t \to \infty} 0
\]

and

\[
\frac{|w_{t+1}|}{|w_1 \ldots w_t|} \xrightarrow{t \to \infty} 0.
\]

Then the sequence $N = w_1 w_2 \ldots$ is normal.
Constructing $\tilde{N}$

- **Corollary:** The sequence

$$S_1 = v_1 v_2 v_2 v_3 v_3 v_3 \ldots$$

obtained by concatenating \( i \) copies of \( v_i \) is normal.
Constructing $L$

- For any word $w \in \{0, 1\}^*$, let

\[ d(w) = \#_0(w) - \#_1(w). \]

Define $L$ to be the language of all words that have as many 0’s as 1’s, i.e.,

\[ L = \{w \in \{0, 1\}^* : d(w) = 0\}. \]

- $L$ is obviously a deterministic one-counter language: store sign and absolute value of $d(\nu)$ ($\nu$ being the scanned prefix of the input) by state and number of stack symbols, respectively.
\( \tilde{N} / L \) is not normal

- Call each \( v_i \) a designated subword. Let \( z_t \) be the prefix of \( \tilde{N} \) that consists of the first \( t \) designated subwords.

- **Proposition:** Among all prefixes \( w \) of \( \tilde{N} \), exactly the prefixes the form

  \[
  z_t = v_1 v_2 v_2 v_3 v_3 v_3 \ldots v_i(t) v_i(t)
  \]

  for any \( t \geq 1 \) satisfy \( d(w) = 0 \), hence are in \( L \).

- Observe: Each designated subword \( v_i \) starts with 0.
**Theorem:** There exists a linear language $L$ and a normal sequence $\hat{N}$ such that the sequence $\hat{N}/L$ selected from $\hat{N}$ by $L$ is infinite and constant.
Constructing $L$

- For any word $w = w(0) \ldots w(n-1)$ of length $n$, let
  \[ w^R = w(n-1) \ldots w(0) \]
  be the mirror word of $w$ and let
  \[ L = \{ww^R : w \in \{0, 1\}^*\} \]
  be the language of palindromes of even length.
- $L$ is linear because it can be generated by a grammar with start symbol $S$ and rules
  \[ S \rightarrow 0S0 \mid 1S1 \mid \lambda. \]
Constructing $\hat{N}$

- $\hat{N}$ is defined in stages $s = 0, 1, \ldots$ where during stage $s$ we specify prefixes $\tilde{z}_s$ and $z_s$ of $N$.
- Start with $\tilde{z}_0 = z_0 = \lambda$ and set

$$\tilde{z}_s = z_{s-1}v_s \ldots v_s \ (2^{s-1} \text{ copies of } v_s),$$

and

$$z_s = \tilde{z}_s \tilde{z}_s^R.$$
Examples of the first $z_i$:

\[ z_1 = v_1 v_R^1, \]
\[ \tilde{z}_2 = v_1 v_R^1 v_2 v_2, \]
\[ z_2 = v_1 v_R^1 v_2 v_2 v_R^2 v_R^1 v_1, \]
\[ \tilde{z}_3 = v_1 v_R^1 v_2 v_2 v_R^2 v_R^1 v_1 v_R^3 v_R^3 v_R^3, \]
\[ z_3 = v_1 v_R^1 v_2 v_2 v_R^2 v_R^2 v_R^1 v_1 v_R^3 v_R^3 v_R^3 v_R^3 v_R^3 v_R^3 v_R^3 v_R^3 v_R^3 v_R^3 v_R^3 v_R^3 v_R^3 v_R^3 v_R^3 v_R^3 v_R^3 v_R^3 v_R^1 v_1 v_R^2 v_2 v_R^2 v_R^2 v_R^2 v_R^2 v_R^1 v_1 v_R^1. \]
Does Not Preserve Normality

- Use Champernowne’s Lemma to show that $\hat{N}$ is normal.

- Proposition: The set of prefixes of $\hat{N}$ that are in $L$ is precisely the set

  \[ \{z_s : s \geq 0\} \]

- It follows that $L$ selects from $\hat{N}$ an infinite subsequence that consists only of 0’s, since any prefix $z_s$ of $\hat{N}$ is followed by the word $v_{s+1}$, where all these words start with 0.
Complexity Issues

• How complex are the counterexamples constructed?
• We want to measure the complexity of the sequence as a language.
• For $\hat{N}$ and $\tilde{N}$, $w \in \hat{N}, \tilde{N}$ can be tested by a nondeterministic linear bounded automaton. Hence $\hat{N}, \tilde{N} \in \text{NSPACE}(O(n))$.
• This means they are both context sensitive.
Complexity Issues

- How complex may these counterexamples be?
- Coding at very distant positions, we can make $\hat{N}, \tilde{N}$ arbitrary complex without destroying normality.
- If we code after a block $z_i$, those places can be ignored by a one counter automaton.