On Universal Types

Gadiel Seroussi
Hewlett-Packard Laboratories
Palo Alto, California, USA

University of Minnesota, September 14, 2004
Types for Parametric Probability Distributions

- \( A = \) finite alphabet, *sequence* or *string* \( x^n = x_1 x_2 \ldots x_n \in A^n \ (n \geq 0) \),
  \( x^k_j = x_j x_{j+1} \ldots x_k \) *sub-string*

- Parametric family \( \mathcal{P} = \{ P_\Theta \} \) of distributions on \( A^n \),
  \[
P_\Theta : A^n \rightarrow [0, 1], \quad \Theta \in \mathcal{D} \subseteq \mathbb{R}^K
  \]

- *Type class* of \( x^n \) with respect to \( \mathcal{P} \):
  \[
  \mathcal{T}_{x^n}^\mathcal{P} = \{ y^n \in A^n : P_\Theta(y^n) = P_\Theta(x^n) \ \forall \Theta \in \mathcal{D} \}
  \]

we say that \( y^n \) and \( x^n \) are *of the same type*
• **Type class** of \( x^n \) with respect to \( \mathcal{P} \):

\[
T_{x^n}^{\mathcal{P}} = \{ y^n \in A^n : P_{\Theta}(y^n) = P_{\Theta}(x^n) \ \forall \Theta \in \mathbb{D} \}
\]

**Example:** \( A = \{0, 1\} \), \( \mathcal{P} = \{\text{Bernoulli}(\theta) \mid 0 \leq \theta \leq 1\} \)

(1-parameter family):

\( x^n \) and \( y^n \) are of the same type \( \Leftrightarrow \) they have the same count of 1's

010101010101 \( \Leftrightarrow \) 000000111111

**Example:** \( k \)-th order binary Markov: \( 2^k \) parameters \( P(x_i = 1 \mid x_{i-k}^{i-1}) \)

... same \( k+1 \)st order joint empirical distributions (with appropriate initial state assumptions)

• The *method of types* is classical in information theory

[Csiszár & Körner '81, Csiszár '98, ...]
The Lempel-Ziv incremental parsing

- Parse $x^n$ into phrases

$$x^n = p_0 \ p_1 \ p_2 \ \cdots \ p_{c-1} \ t_x$$

- $p_0 = \lambda$, the null string
- $p_i$, $i > 0$, is the *shortest substring* of $x^n$, starting at the position following $p_{i-1}$, such that $p_i \neq p_j$, $\forall j < i$ (all $p_i$'s are different)
- every phrase $p_i$, except $p_0 = \lambda$, is an extension of another phrase
- $t_x$ is the remaining *tail* of $x^n$ where parsing is truncated due to the end of the sequence (must be equal to one of the $p_i$'s)

- **LZ78 universal data compression algorithm** [Ziv & Lempel, 1978] has code length $L_{LZ} = c \log c + O(c)$

Example:

$$x^8 = 1 \ 0 \ 1 \ 0 \ 1 \ 1 \ 0 \ 0$$

$$\rightarrow \ \lambda, \ 1, \ 0, \ 1 \ 0, \ 1 \ 1, \ 0 \ 0$$

$c = 6$, $t_x = \lambda$

![Tree representation](image-url)
Universal type classes

- $T_{x^n} = \{p_0, p_1, \ldots, p_{c-1}\}$ set of phrases of $x^n$

- Universal (LZ) type class of $x^n$:

  $$\mathcal{U}_{x^n} = \{ y^n \in A^n : T_{y^n} = T_{x^n} \}$$

  all strings of length $n$ that produce the same set of phrases as $x^n$

- same phrases, generally in different order

- phrase permutation must respect prefix relation

- Example:

  $$x^8 = 10101100 \rightarrow 1, 0, 10, 11, 00$$

  $$\rightarrow 0, 00, 1, 11, 10 \rightarrow 00011110 = y^8$$
Why are universal type classes interesting?

• $N(u^k, v^m) \overset{\Delta}{=} \text{number of (possibly overlapping) occurrences of } u^k \text{ in } v^m$.

\[
\hat{P}_{x^n}^{(k)}(u^k) \overset{\Delta}{=} \frac{N(u^k, x^n)}{(n - k + 1)}, \quad u^k \in A^k
\]

Empirical joint distribution of order $k \leq n$

Theorem. Let $x^n$ be an arbitrary sequence of length $n$, and $k$ a fixed positive integer. If $y^n \in U_{x^n}$, then, for all $u^k \in A^k$,

\[
\hat{P}_{x^n}^{(k)}(u^k) - \hat{P}_{y^n}^{(k)}(u^k) \to 0 \quad \text{as} \quad n \to \infty.
\]

(empirical distributions of any order $k$ converge in the variational sense)

• Asymptotics: an infinite sequence of sequences $x^n$, not necessarily prefix-related

• Proof arguments:
  - any $u^k$ is either contained in a phrase, or spans a phrase boundary ($O(kc)$ places), or contained in the tail ($\leq |t_x|$ places)
  - well known LZ properties: $c \leq n/\left(\log n - o(\log n)\right)$, $|p_i| \leq \sqrt{2n} - o(\sqrt{n})$

• Convergence: $O(1/\log n)$
Finite memory probability assignments

- A \( k \)-th order (finite-memory) probability assignment on \( A^n \) is defined by
  - a set of conditional distributions \( Q_k(u_{k+1}|u^k), \ u^{k+1} \in A^{k+1} \)
  - a distribution \( Q_k(u^k) \) on the initial state
  - so that \( Q_k(x^n) = Q_k(x^k) \prod_{i=k+1}^n Q_k(x_i|x_{i-1}^{i-k}) \)

- In particular, \( Q_k \) could be defined by the \( k \)-th order approximation of an ergodic measure

**Corollary.** Let \( x^n \) and \( y^n \) be sequences of the same universal type. Then, for any nonnegative integer \( k \) and any \( k \)-th order probability assignment \( Q_k \) s.t. \( Q_k(x^n), Q_k(y^n) \neq 0 \), we have

\[
\frac{1}{n} \log \frac{Q_k(x^n)}{Q_k(y^n)} \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty
\]
**Conventional vs. Universal types**

### Sequences of the same ...

<table>
<thead>
<tr>
<th>conventional type $T_x^n$</th>
<th>universal type $U_x^n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>have <em>identical</em> empirical distributions of a <em>given</em> structure</td>
<td>have <em>converging</em> empirical distributions of <em>any</em> finite order</td>
</tr>
<tr>
<td>are assigned <em>identical</em> probabilities by assignments from a <em>given</em> parametric class</td>
<td>are assigned probabilities whose normalized logs <em>converge</em>, for finite-memory assignments of <em>any</em> finite order</td>
</tr>
</tbody>
</table>

*degree of “indistinguishability” vs. universality*
Other things we like to do with type classes

- Ask how many sequences in the type class of \( x^n \)
  (conventional: \( \approx 2^n \hat{H}(x^n) + o(n) \) )

- Ask how many types for given seq. length \( n \)
  (conventional: \( \approx n^K \) )

- Draw a sequence at random with uniform probability from the class

- Enumerate the type class

- Use the class for (enumerative) coding and universal simulation
The tree is the type (almost)

- **Assumptions**
  - For simplicity, \( A = \{0, 1\} \) (everything carries to \(|A| > 2\))
  - We ignore effects of tails (nuisance, does not affect main results)
  - When in doubt, \( \log \) is to base 2

- **Parsing tree**
  - \( c = \text{number of nodes} \)
  - \( n = \sum_i |p_i| = \text{path length of } T \)
  - \( \ell = \text{number of leaves} \)
  - \( h = \text{height} = \text{length of longest path} \)

- All sequences in \( U^n_x \) yield the same tree: *the type is* \( T \) (or \( (T, n) \))

\[
\begin{align*}
T & \quad \lambda \\
0 & \quad 0 \\
1 & \quad 1 \\
00 & \quad 10 \\
11 & \\
\end{align*}
\]

\( c = 6, \ n = 8, \ \ell = 3, \ h = 2 \)

*a combinatorial characterization, not explicitly based on counts*
Size of the universal type class

• Phrases coming from $T_0$ and $T_1$ do not have prefix relation constraints, and can be interleaved arbitrarily

Let $\mathcal{U}_T = \text{type class associated with a tree } T$. Then

$$|\mathcal{U}_T| = |\mathcal{U}_{T_0}| \cdot |\mathcal{U}_{T_1}| \cdot \binom{c-1}{c_0}$$

• note: $c - 1 = c_0 + c_1$

• cases where a sub-tree is missing are easily accounted for (deficient trees)

Let $c_{p_i} = \text{size of sub-tree rooted at } p_i$. Then

$$|\mathcal{U}_T| = \frac{(c-1)!}{\prod_{i>0} c_{p_i}}$$
Size of the universal type class—example

\[ x^8 = 10101100, \quad c = 6 \]
\[ c_0 = 2, \quad c_{00} = 1, \quad c_1 = 3, \quad c_{10} = c_{11} = 1 \]
\[ |\mathcal{U}_T| = \frac{5!}{2 \cdot 3} = 20 \]

0, 00, 1, 10, 11 0, 00, 1, 11, 10
0, 1, 00, 10, 11 0, 1, 00, 11, 10
0, 1, 10, 00, 11 0, 1, 11, 00, 10
0, 1, 10, 11, 00 0, 1, 11, 10, 00
1, 0, 00, 10, 11 1, 0, 00, 11, 10
1, 0, 10, 00, 11 1, 0, 11, 00, 10
1, 0, 10, 11, 00 1, 0, 11, 10, 00
1, 10, 0, 00, 11 1, 11, 0, 11, 10
1, 10, 0, 11, 00 1, 11, 0, 10, 00
1, 10, 11, 0, 00 1, 10, 11, 0, 00
Asymptotic size of a universal type class

Theorem. The size of $\mathcal{U}_{x^n}$ satisfies

$$\frac{1}{n} \left( \log |\mathcal{U}_{x^n}| - \mathcal{L}_{LZ}(x^n) \right) \to 0 \quad \text{as} \quad n \to \infty$$

Analogous to conventional types:

$$\frac{1}{n} \left( \log |\mathcal{T}_{x^n}| - \hat{H}(x^n) \right) \to 0 \quad \text{as} \quad n \to \infty$$

with $\mathcal{L}_{LZ}(x^n) \approx c \log c$ in lieu of the empirical entropy $\hat{H}(x^n)$

(entropy of ML distribution)
Asymptotic size of a universal type class (zoom-in)

Theorem. The size of $\mathcal{U}_T$ satisfies

$$(1 - \beta - o(1)) \ c \log c \leq \log |\mathcal{U}_T| \leq (1 - \eta - o(1)) \ c \log c,$$

where $0 \leq \beta, \eta \leq 1$, and

- $\beta$ is bounded away from $0$ iff $\frac{\log n}{\log c}$ is bounded away from $1$,
- $\eta$ is bounded away from $0$ iff $\frac{\log n}{\log c} \to 2$.

Notes:

- we have $1 \leq \frac{\log n}{\log c} \leq 2$ (up to lower order terms)
- $\frac{\log n}{\log c}$ bounded away from $1 \Rightarrow c \log c = o(n) \Rightarrow \frac{1}{n} \mathcal{L}_{LZ} \to 0$
- more detail: $\beta = (1 - \ell) \left( \frac{\log n}{\log c} - 1 \right)$, $\eta = \frac{(h+1) \log(h+1)}{c \log c}$
Asymptotic size of the universal type: proof outline

- \( |\mathcal{U}_T| = \frac{(c - 1)!}{\prod_{i>0} c_{p_i}} \triangleq \frac{(c - 1)!}{D} \) need to estimate \( D = \prod_{i>0} c_{p_i} \) (hard?)

- let’s estimate the sum instead

\[
S = \sum_{i>0} c_{p_i} = n \quad \text{easy!!!}
\]

- if \( p_i \) is a leaf, then \( c_{p_i} = 1 \). Rewrite sum and product:

\[
S' = \sum_{c_{p_i}>1} c_{p_i} = n - \ell ; \quad D = \prod_{i>0} c_{p_i} = \prod_{c_{p_i}>1} c_{p_i}
\]

- \( D = \text{product of positive integers whose sum is known} \)

\[
\Rightarrow D \leq \left( \frac{n - \ell}{c - \ell - 1} \right)^{c-\ell-1} \quad \text{the rest follows}
\]
Examples

**Skinny full tree**

\[ c = 2k + 1, \]
\[ n = k(k + 1) = \frac{c^2 - 1}{4}, \]
\[ \frac{\log c}{\log n} \rightarrow \frac{1}{2}, \]
\[ \log |U_T| = \log \left(2^k k!\right) = \frac{1}{2} c \log c + O(c) \]

**Fat tree**

\[ c = 2^{m+1} - 2 \]
\[ n = \sum_{i=1}^{m} i2^i = (m - 1)2^{m+1} + 2 \]
\[ \frac{\log c}{\log n} \rightarrow 1, \]
\[ \log |U_T| = c \log c + O(c) \]

0 \leq (1 - \beta) \leq \frac{1}{2} \text{ attained with deficient trees, e.g.:}

totally skinny \quad \bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \Rightarrow \log |U_T| = 0
Coding with the universal type

- **Enumerative coding [Cover 1973]:** encode $x^n$ by describing its type + index of the sequence within the type

- For universal types
  - Description of the parsing tree: $2c + o(c)$ bits
  - Index within the type: $\log |\mathcal{U}_{x^n}|$
  - In some cases, total description length can be shorter than LZ78 code length, e.g. for skinny full tree

\[
\text{description length} = \frac{1}{2} c \log c + O(c)
\]

possible only when $\frac{1}{n} \mathcal{L}_{\text{LZ}} \to 0$
How many universal types?

- Number of types for sequences of length $n$
  
  $= \text{number of LZ78 parsing dictionaries}$
  
  $= \text{number of binary trees of given path length } n$

- **Known (hard?) open combinatorial problem**
  
  - $b_{c,n} = \text{number of trees with } c \text{ nodes and path length } n = n$
  
  - $B(w, z) = \sum_{c,n} b_{c,n} w^n z^c$ generating function satisfies
    
    $$zB(w, wz)^2 = B(w, z) - 1 \quad (\text{we need } B(w, 1))$$

  - “Airy phenomenon”, Brownian motion, analysis of quicksort, ...
How many universal types?

- Using information-theoretic/combinatorial arguments

**Theorem.** Let

\[ \tau_n = \log \left( \text{number of types for sequences of length } n \right). \]

Then,

\[ \tau_n = \frac{2n}{\log n} \left( 1 + o(1) \right) \]

Sloane's OEIS #A095830—Number of binary trees for given path length

G. Seroussi, “On the number of t-ary trees with a given path length,”

How many universal types?

- Using information-theoretic/combinatorial arguments

**Theorem.** Let $t = |A|$ and

$$\tau_n = \log_t (\text{number of types for sequences of length } n).$$

Then,

$$\tau_n = \frac{t H_2(t^{-1}) n}{\log n} \left(1 + o(1)\right)$$

Sloane's OEIS #A095830—Number of binary trees for given path length

G. Seroussi, “On the number of $t$-ary trees with a given path length,”

**Bounds**

**Upper bound.**

\[ \tau_n \leq \frac{2n}{\log n} \left(1 + o(1)\right) \]

- a binary tree with \( c \) nodes can be encoded using \( 2c \) bits
- a tree with path length \( n \) has at most \( c = \frac{n}{\log n - O(\log \log n)} \) nodes
- \[ \Rightarrow \frac{2n}{\log n - O(\log \log n)} \] bits are sufficient to (losslessly) encode all trees of path length \( n \)

**Lower bound.**

\[ \tau_n \geq \frac{2n}{\log n} \left(1 - o(1)\right) \]

- proof by building a large class of trees of the same path length

\[ 2^{m-1} \] nodes at depth \( m-1 \)

\( \ell - 1 \) different sub-trees of size \( \approx \frac{m}{2} \)

(\( \ell - 1 \))! permutations \( \Rightarrow \) (\( \ell - 1 \))! different trees of same path length
• Mark phrases as *unused* or *used* (initially all *unused* except the root \( \lambda \))
• Pick next phrase by starting from the root and following branches to an unused phrase
• When at a bifurcation, choose branch randomly with probabilities proportional to number of unused nodes in sub-tree

**Example:**

At root: choose branch with

\[ \text{Prob}(0) = \frac{2}{5}, \quad \text{Prob}(1) = \frac{3}{5} \]

Choose \( y = 1 \)

\[ P = \frac{3}{5} \]
Drawing a random sequence from the universal type

- Mark phrases as *unused* or *used* (initially all *unused* except the root $\lambda$)
- Pick next phrase by starting from the root and following branches to an unused phrase
- When at a bifurcation, choose branch randomly with probabilities proportional to number of unused nodes in sub-tree

**Example:**

At root: choose branch with

$$\text{Prob}(0) = \frac{1}{2}, \quad \text{Prob}(1) = \frac{1}{2}$$

Choose $\frac{1}{2}$

$$P = \frac{3}{5} \cdot \frac{1}{2}$$

reached a used node, keep going
Drawing a random sequence from the universal type

- Mark phrases as *unused* or *used* (initially all *unused* except the root $\lambda$)
- Pick next phrase by starting from the root and following branches to an unused phrase
- When at a bifurcation, choose branch randomly with probabilities proportional to number of unused nodes in sub-tree

**Example:**

At '1': choose branch with

$\text{Prob}(0) = \frac{1}{2}$, $\text{Prob}(1) = \frac{1}{2}$

Choose 0

$P = \frac{3}{5} \cdot \frac{1}{2} \cdot \frac{1}{2}$

$y = 110$
Drawing a random sequence from the universal type

- Mark phrases as unused or used (initially all unused except the root \( \lambda \))
- Pick next phrase by starting from the root and following branches to an unused phrase
- When at a bifurcation, choose branch randomly with probabilities proportional to number of unused nodes in sub-tree

Example:

At root: choose branch with 
\[
\text{Prob}(0) = \frac{2}{3}, \quad \text{Prob}(1) = \frac{1}{3}
\]

Choose 0
\[
P = \frac{3}{5} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{2}{3}
\]
\[
y = 1100
\]
Drawing a random sequence from the universal type

- Mark phrases as unused or used (initially all unused except the root $\lambda$)
- Pick next phrase by starting from the root and following branches to an unused phrase
- When at a bifurcation, choose branch randomly with probabilities proportional to number of unused nodes in sub-tree

**Example:**

At root: choose branch with $\text{Prob}(0) = \frac{1}{2}$, $\text{Prob}(1) = \frac{1}{2}$

Choose 0 (rest is deterministic)

$$P = \frac{3}{5} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{20} = \frac{1}{|U_T|}$$

$y = 1 \ 1 \ 0 \ 0 \ 0 \ 0 \ 1 \ 1$
Proposition. The procedure outlined

(i) produces a sequence \( y^n \in U_x^n \) with uniform probability

(ii) requires an average of \( \log |U_x^n| + O(1) \) random bits to do so (i.e., it is “entropy-efficient”)

(ii) is accomplished by letting an arithmetic decoder, fed with a Bernoulli(\( \frac{1}{2} \)) sequence and the sequence of bifurcation probabilities, generate the random choices called for by the procedure [Han-Hoshi’97]
Universal simulation of individual sequences

- Given an *individual sequence* $x^n$, we wish
  1. to produce a sequence $y^n$ that is statistically “similar” to $x^n$
  2. to let $y^n$ have as much uncertainty (entropy) as possible given the previous requirement

- Proposed solution: choose $y^n$ randomly and uniformly from $U_{x^n}$
  1. “similarity” = convergence in variation of empirical distributions of any order
  2. asymptotically optimal in the following sense: any simulation scheme satisfying (1) must have

$$H(y^n) \leq \log |U_{x^n}| + \Delta \quad \forall \Delta > 0 \quad \text{and almost all } x^n$$

(doubly-universal analogue of [Merhav-Weinberger’04])
Example: Simulation of a binary texture

Simulation of a binary texture
(512 \times 512 \text{ out of } 1024 \times 1024, \text{ scanned with Peano scan})
Example: Simulation of a binary texture

Simulation of a binary texture
(512 × 512 out of 1024 × 1024, scanned with Peano scan)
Additional results and research questions

- Explicit enumeration of the universal type. Given $x^n$, efficiently implement the map

$$\{1, 2, 3, \ldots, |U_{x^n}|\} \leftrightarrow U_{x^n}$$

- provides an alternative way to draw uniformly from the type class
- trade-off: fixed number of random bits with approximate uniform probability vs. variable number of random bits with exact uniform probability
- allows implementation of enumerative code

- Ongoing research: study universal types for other universal compression algorithms (e.g., Rissanen’s Context)
  - no problems in principle, but combinatorial analysis seems more challenging than LZ