A Tree Sampler for Bounded Context-Free Languages

Breandan Considine

Main Idea

- If you can count it, you can sample it! Analytic combinatorics
- We implement a bijection between binary trees in bounded CFLs
- Allows for parallelizable replacement-free sampling from BCFLs

Algebraic Parsing

Given a CFG $\mathcal{G} \coloneqq \langle V, \Sigma, P, S\rangle$ in Chomsky Normal Form (CNF), we may construct a recognizer $R_{\mathcal{G}} : \Sigma^n \to \mathbb{B}$ for strings $\sigma : \Sigma^n$ as follows. Let $\in^V$ be our domain, where 0 is $\varnothing$, $\oplus$ is $\cup$, and $\otimes$ be defined as:

$$s_1 \otimes s_2 \coloneqq \{C \mid \langle A, B\rangle \in s_1 \times s_2, (C \to AB) \in P\}$$

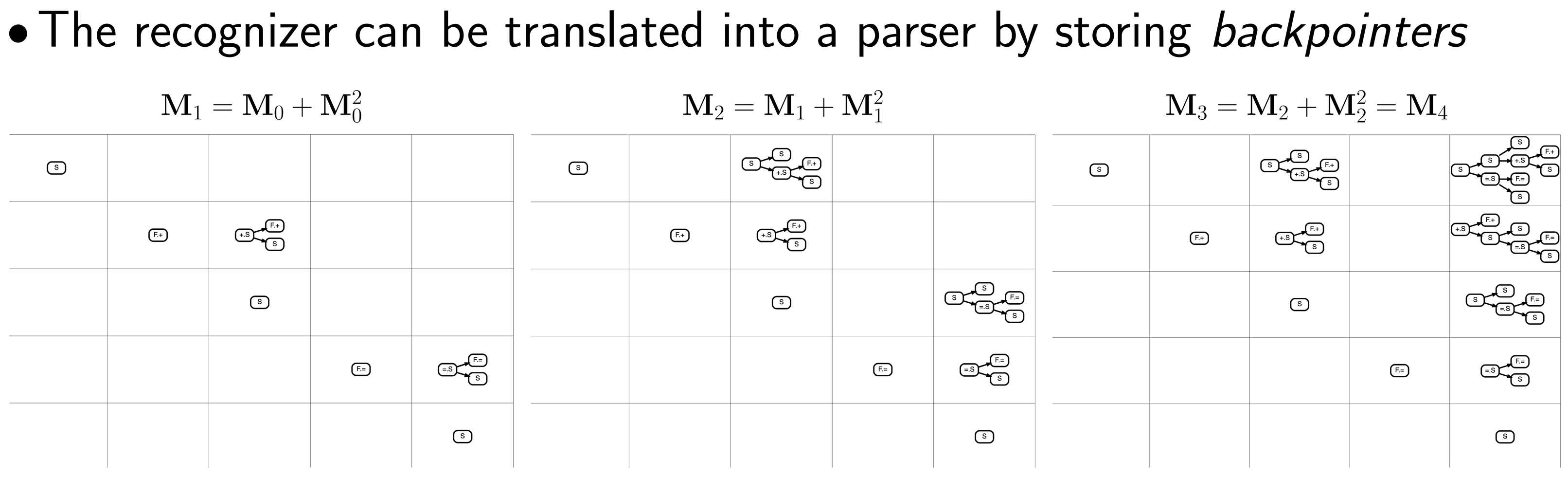
Initializing $\mathbf{M}_0[i,j](\mathcal{G},\sigma) \coloneqq \{A \mid i+1 = j, (A \to \sigma_i) \in P\}$ and searching for the least solution to $\mathbf{M} = \mathbf{M} + \mathbf{M}^2$, will produce a fixedpoint $\mathbf{M}^*$:

$$M^0 \coloneqq \begin{pmatrix} \varnothing & \hat\sigma_1 & \varnothing & \varnothing \\ & & & \varnothing \\ & & & \hat\sigma_n \\ \varnothing & & & \varnothing \end{pmatrix} \Rightarrow M_\infty = \begin{pmatrix} \varnothing & \hat\sigma_1 & \Lambda & \Lambda^*_\sigma \\ & & & \Lambda \\ & & & \hat\sigma_n \\ \varnothing & & & \varnothing \end{pmatrix}$$

Valiant (1975) shows that $\sigma \in \mathcal{L}(\mathcal{G})$ iff $S \in \mathcal{T}$, i.e., $\mathbb{1}_{\mathcal{T}}(S) \iff \mathbb{1}_{\mathcal{L}(\mathcal{G})}(\sigma)$.

Parsing Dynamics

- The matrix $\mathbf{M}_0$ is strictly upper triangular, i.e., nilpotent of degree n
- The recognizer can be translated into a parser by storing *backpointers*



- If we had a way to solve for $\mathbf{M} = \mathbf{M} + \mathbf{M}^2$ directly, power iteration would be unnecessary and we could solve for $\mathbf{M} = \mathbf{M}^2$ above the superdiagonal...

Binarized CFL Sketching

- CYK parser can be lowered onto a Boolean tensor $\mathbb{B}^{n\times n\times|V|}$ (Valiant, 1975)
- Binarized CYK parser can be compiled to SAT to solve for $\mathbf{M}^*$ directly
- Enables sketch-based synthesis in either σ or $\mathcal{G}$: just use variables for holes!
- We simply encode the characteristic function, i.e. $\mathbb{1}_{\subseteq V} : V \to \mathbb{B}^{|V|}$
- $\oplus, \otimes$ are defined as $\boxplus, \boxtimes$, so that the following diagram commutes:

$$\begin{tikzcd} 2^V \times 2^V \arrow[r, "\oplus/\otimes"] \arrow[d, shift left, "\mathbb{1}^2"] & 2^V \arrow[d, shift left, "\mathbb{1}"] \\ \mathbb{B}^{|V|} \times \mathbb{B}^{|V|} \arrow[r, "\boxplus/\boxtimes"'] \arrow[u, shift left, "\mathbb{1}^{-2}"] & \mathbb{B}^{|V|} \arrow[u, shift left, "\mathbb{1}^{-1}"] \end{tikzcd}$$

- These operators can be lifted into matrices and tensors in the usual way

Method

We define an algebraic data type $\mathbb{T}_3 = (V \cup \Sigma) \rightharpoonup \mathbb{T}_2$ where $\mathbb{T}_2 = (V \cup \Sigma) \times (\mathbb{N} \rightharpoonup \mathbb{T}_2 \times \mathbb{T}_2)$[a]. Morally, we can think of $\mathbb{T}_2$ as an implicit set of possible trees sharing the same root, and $\mathbb{T}_3$ as a dictionary of possible $\mathbb{T}_2$ values indexed by possible roots, given by a specific CFG under a finite-length porous string. We construct $\hat\sigma_r = \Lambda(\sigma_r)$ as follows:

$$\Lambda(s : \underline{\Sigma}) \mapsto \begin{cases} \bigoplus_{s\in\Sigma} \Lambda(s) & \text{if } s \text{ is a hole,} \\ \big\{\mathbb{T}_2\big(w, \big[\langle \mathbb{T}_2(s), \mathbb{T}_2(\varepsilon)\rangle\big]\big) \mid (w \to s) \in P\big\} & \text{otherwise.} \end{cases}$$

We then compute the fixpoint M_∞ by redefining the operations $\oplus, \otimes : \mathbb{T}_3 \times \mathbb{T}_3 \to \mathbb{T}_3$ as follows:

$$X \oplus Z \mapsto \bigcup_{k\in\pi_1(X\cup Z)} \Big\{k \Rightarrow \mathbb{T}_2(k, x \cup z) \mid x \in \pi_2(X \circ k), z \in \pi_2(Z \circ k)\Big\}$$

$$X \otimes Z \mapsto \bigoplus_{(w\to xz)\in P} \Big\{\mathbb{T}_2\Big(w, \big[\langle X \circ x, Z \circ z\rangle\big]\Big) \mid x \in \pi_1(X), z \in \pi_1(Z)\Big\}$$

These operators group subtrees by their root nonterminal, then aggregate their children. Each Λ now becomes a dictionary indexed by the root nonterminal, which can be sampled by obtaining $(\Lambda^*_\sigma \circ S) : \mathbb{T}_2$, then recursively choosing twins.

[a] Given a $T : \mathbb{T}_2$, we may also refer to $\pi_1(T), \pi_2(T)$ as $\texttt{root}(T)$ and $\texttt{children}(T)$ respectively, where children are pairs of conjoined twins.

A Pairing Function for BTrees

Let $\mathbf{M}$: GF$(2^{n\times n})$ be a square matrix $\mathbf{M}^0_{r,c} = P_c$ if $r = 0$ else $\mathbb{1}[c = r-1]$, where P is a feedback polynomial with coefficients $P_{1\ldots n}$ and $\oplus \coloneqq \veebar, \otimes \coloneqq \wedge$:

$$\varphi(T : \mathbb{T}_2, i : \mathbb{Z}_{|T|}) \mapsto \begin{cases} \Big\langle \texttt{BTree}\big(\texttt{root}(T)\big), i\Big\rangle & \text{if } T \text{ is a leaf,} \\ \text{Let } b = |\texttt{children}(T)|, & \\ \quad q_1, r = \Big\langle \lfloor \tfrac{i}{b} \rfloor, i \pmod b\Big\rangle, & \\ \quad lb, rb = \texttt{children}[r], & \\ \quad T_1, q_2 = \varphi(lb, q_1), & \\ \quad T_2, q_3 = \varphi(rb, q_2) \text{ in} & \\ \Big\langle \texttt{BTree}\big(\texttt{root}(T), T_1, T_2\big), q_3\Big\rangle & \text{otherwise.} \end{cases}$$

Selecting any $V \neq \mathbf{0}$ and coefficients P_j from a known *primitive polynomial*, then powering the matrix $\mathbf{M}$ generates an ergodic sequence over GF(2^n):

Tidyparse IDE Plugin