

# Backpropagation of Syntax Errors in Context-Sensitive Languages

Breandan Considine, Jin Guo, Xujie Si

## Main Idea

- Matrices over  $\mathbb{Z}_2^n$  are useful structures for studying finite state machines
- The operators  $\{\text{XOR}, \wedge, \top\}$  are *functionally complete* logical primitives
- We use them to implement probabilistic context-sensitive program repair

## Algebraic Parsing

Given a CFG,  $\mathcal{G}' : \langle \Sigma, V, P, S \rangle$  in Chomsky Normal Form (CNF), we can define a *recognizer*,  $R : \mathcal{G}' \rightarrow \Sigma^n \rightarrow \mathbb{B}$  for bounded strings  $\sigma : \Sigma^n$  using the following construction. Let  $2^V$  be our domain, 0 be  $\emptyset$ ,  $\oplus$  be  $\cup$ , and  $\otimes$ :

$$X \otimes Z := \{ w \mid \langle x, z \rangle \in X \times Z, (w \rightarrow xz) \in P \}$$

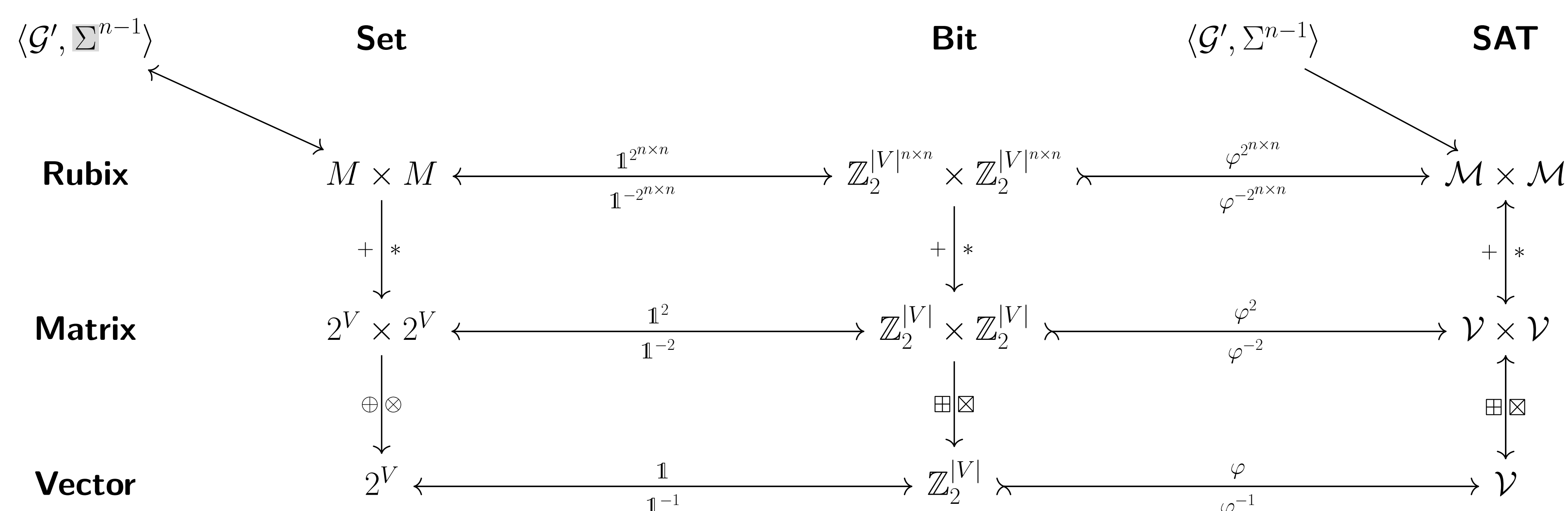
Valiant (1975) shows that if we let  $\sigma_r^\dagger := \{w \mid (w \rightarrow \sigma_r^\dagger) \in P\}$ , initialize the matrix  $M_{r+1=c}^0(\mathcal{G}', e) := \sigma_r^\dagger$  and solve for its fixpoint  $M^* = M + M^2$ ,

$$M^0 := \begin{pmatrix} \emptyset & \sigma_1^\dagger & \emptyset & \dots & \emptyset \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \emptyset & \dots & \emptyset & \dots & \sigma_n^\dagger \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \emptyset & \dots & \emptyset & \dots & \emptyset \end{pmatrix} \Rightarrow M^* = \begin{pmatrix} \emptyset & \sigma_1^\dagger & \Lambda & \dots & \Lambda_\sigma^* \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \emptyset & \dots & \emptyset & \dots & \sigma_n^\dagger \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \emptyset & \dots & \emptyset & \dots & \emptyset \end{pmatrix}$$

the recognizer is then defined as:  $R(\mathcal{G}', \sigma) := S \in \Lambda_\sigma^*? \iff \sigma \in \mathcal{L}(\mathcal{G})?$

## Galois Connection

- CYK parsers can be lowered onto  $\mathbb{Z}_2^{|V| \times n \times n}$  or  $\mathcal{M} : (\mathbb{Z}_2^{|V|} \rightarrow \mathbb{Z}_2)^{|V| \times n \times n}$
- $\mathcal{M}^*$  can be solved for directly using Gaussian elimination or XOR-SAT
- Enables sketch-based synthesis in  $\sigma$  or  $\mathcal{G}$ : just use variables for holes!
- We can encode using the characteristic function, i.e.,  $\mathbb{1}_{\subseteq V} : 2^V \rightarrow \mathbb{Z}_2^{|V|}$
- $\oplus, \otimes$  are defined as  $\boxplus, \boxtimes$ , so that the following diagram commutes:



## Brozowski's Derivative

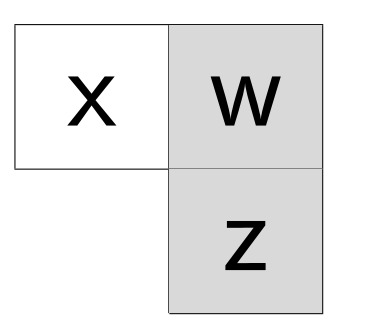
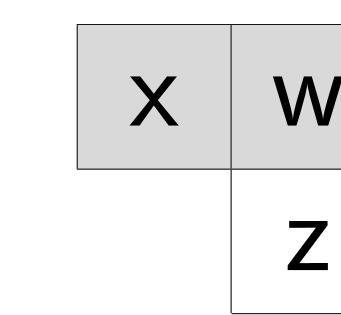
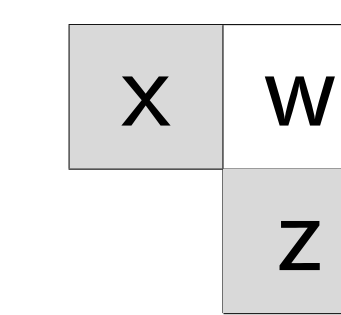
Valiant's  $\otimes$  operator, which unifies known factors in a binary CFG, implies a left- and right-quotient, which yield the set of nonterminal forests that may appear to either side of a known factor and its corresponding root.

Valiant's  $\otimes$

Left Quotient

Right Quotient

$$x \otimes z = \{ w \mid (w \rightarrow xz) \} \quad \frac{\partial f}{\partial x} = \{ z \mid (w \rightarrow xz) \} \quad \frac{\partial f}{\partial z} = \{ x \mid (w \rightarrow xz) \}$$



The left quotient coincides with Brzozowski's derivative (1964) over regular languages, here lifted into the context-sensitive setting (our work).

## Context Sensitivity

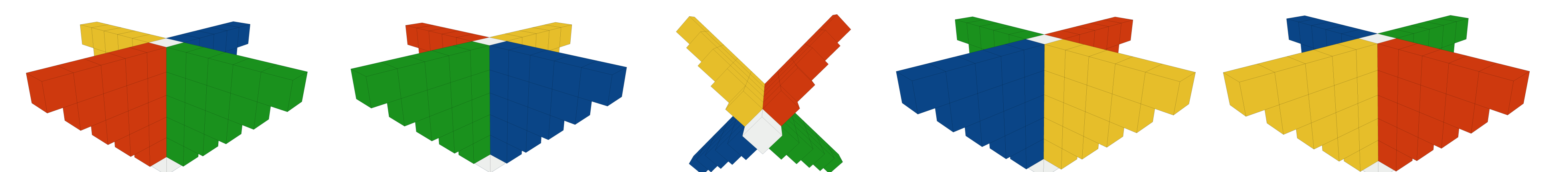
It is well-known that the family of CFLs is not closed under intersection. For example, consider  $\mathcal{L}_\cap := \mathcal{L}(\mathcal{G}_1) \cap \mathcal{L}(\mathcal{G}_1)$  defined in the following way:

$$P_1 := \{ S \rightarrow LR, L \rightarrow ab \mid aLb, R \rightarrow c \mid cR \}$$

$$P_2 := \{ S \rightarrow LR, R \rightarrow bc \mid bRc, L \rightarrow a \mid aL \}$$

$\mathcal{L}_\cap$  is equivalent to the language  $\{ a^d b^d c^d \mid d > 0 \}$ , which is not a CFL. We can encode  $\bigcap_{i=1}^c \mathcal{L}(\mathcal{G}_i)$  as a polygonal prism with upper-triangular matrices adjoined to each rectangular face. Specifically, we intersect all terminals  $\Sigma_\cap := \bigcap_{i=1}^c \Sigma_i$ , then for each  $t \in \Sigma_\cap$ , construct an equivalence class  $E(t, \mathcal{G}_i) = \{w_i \mid (w_i \rightarrow t) \in P_i\}$  and glue them together at each  $\sigma_i$ :

$$\bigwedge_{t \in \Sigma_\cap} \bigwedge_{j=1}^{c-1} \bigwedge_{i=1}^{|\sigma|} E(t, \mathcal{G}_j) \equiv_{\sigma_i} E(t, \mathcal{G}_{j+1})$$



Orientations of a  $\bigcap_{i=1}^4 \mathcal{L}(\mathcal{G}_i) \cap \Sigma^6$  configuration, reprojected into 2-space.

As  $c \rightarrow \infty$ , this shape approximates a circular cone whose symmetric axis intersects orthonormal CNF unit productions  $w_i \rightarrow t$ , with  $S_i \in \Lambda_\sigma^*$  encoded by bitvectors on the base perimeter. Equations of this form are equiexpressive with the family of CSLs realizable by finite CFL intersection.

